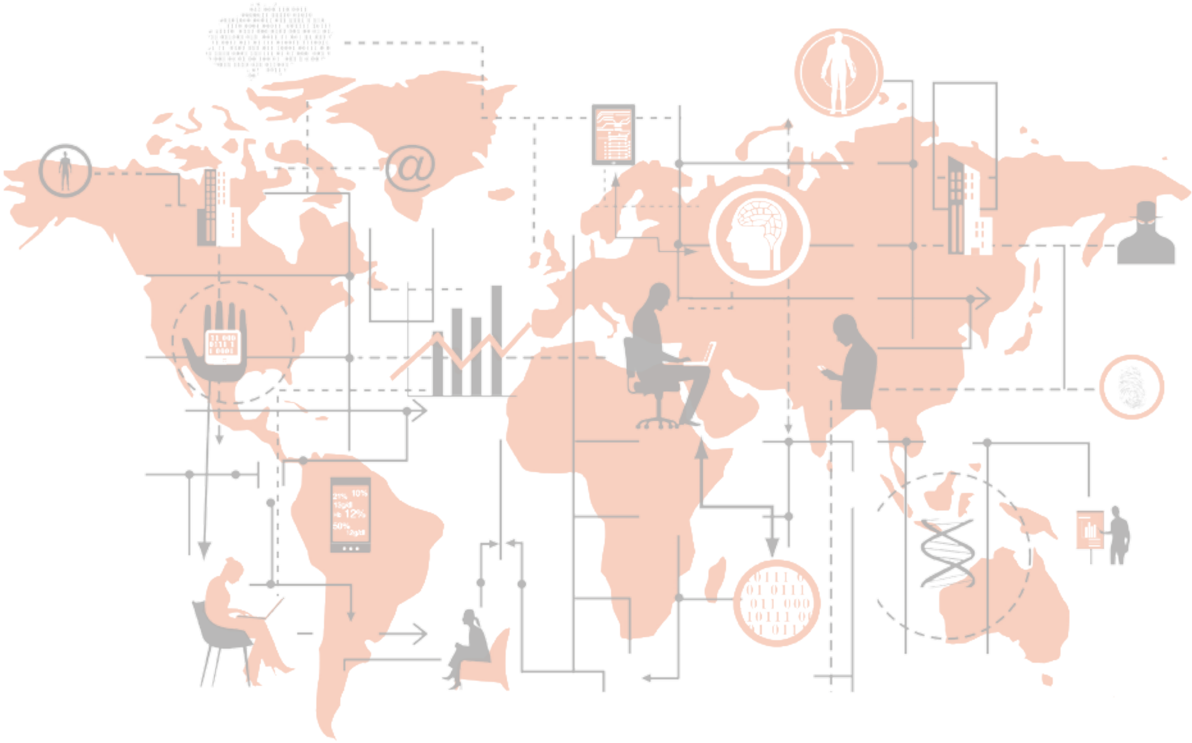
NO MORE SECRETS
with Big Data Analytics

Jaap Bloem, Menno van Doorn, Sander Duivestein, Thomas van Manen, Erik van Ommeren, Sandeep Sachdeva

SOGETI

No More Secrets with Big Data Analytics



Jaap Bloem, Menno van Doorn, Sander Duivestein,
Thomas van Manen, Erik van Ommeren, Sandeep Sachdeva

SOGETI

VINT | Vision • Inspiration • Navigation • Trends

Book production LINE UP boek en media bv, Groningen

Table of Contents

Foreword
from the VINT Board of Recommendation

Daily blog posts about Big Data and four reports on this new field preceded this book. In *No More Secrets* the authors combine their findings. Their focus is the new opportunities and challenges offered by Big Data Analytics. The insight that data is the new oil catches on and Big Data has become a prominent topic of discussion in boardrooms.

Big Data Analytics is becoming good practice in ever more domains: this fact lingers most after reading *No More Secrets*. Some organizations are well under way, others are accelerating or have started experiments. The consequence of this burgeoning Big Data success now starts to play out: secrets, big and small, have no future. Searching for oil or the way DNA rules our health used to be covered in mystery, as was consumer behavior. Thanks to advanced Big Data Analytics one secret after another is now being unraveled.

Rooted in data-intensive science, Big Data Analytics now is being deployed everywhere. The economy and everyday life are full of guiding examples. Big Data techniques help the discovery of buying patterns and the detection of fraud. The cases in this book on predicting human behavior from all digital traces we leave behind, incite reflection and inspiration. Based on real-time tweets the inflation rate can be determined: one of those remarkable things that Big Data Analytics has made possible.

No More Secrets may provide the basis for updating or refining your understanding of Big Data Analytics and for exploring new ground. The first Part sheds light on the Big Data phenomenon in general. Part II presents ample suggestions for determining your specific Big Data potential. These you can readily apply to gain insight in what exactly makes your customers tick, the topic of Part III. The triad of privacy, technology and the law concludes the book, while the introductory section after this foreword sketches out the actual context of Big Data along five leading themes.

The warning that Big Data Analytics is not a technology toy but an integral part of strategy, marketing, HR or R&D may already be familiar. The authors go even further and urge you to create “magic moments,” since abandoning conventional thinking and patterns will cost at least as much energy as getting your unstructured data analysis right.

The answer to the fundamental question of what it means to live in a world where secrets of consumers and citizens, and mysteries of our Earth and life itself have been unraveled, became top of mind when Edward Snowden had his revelations published about the digital practices of secret services. These Big Data Analytics fully justify the title *No More Secrets*. Suddenly, Big Data itself had lost its final mystery as this Big Brotherhood of competing and cooperating agencies and companies for the sake of safety was exposed.

In this book the authors decided to focus soleley on the fair commercial use and business impact of Big Data Analytics. We sincerely hope that the insights and ideas in *No More Secrets* will contribute to successful new strategies, innovative implementations, and faster and better business decisions.

Board of Recommendation Sogeti VINT

Introduction
The Future of Big Data Analytics

While working on our Big Data reports and on this book, we met many marketeers and geeks, CIOs and managers, lawyers, activists, forerunners, followers and laggards in the emerging field of Big Data Analytics. Five main themes surfaced: acceleration, transformation, data ownership, privacy and Edward Snowden.

1 Big Data into gear

Acceleration is an important one that Big Data projects have in common. The V of Velocity turns out to be the most popular of the defining Big Data triad of Volume, Variety and Velocity. Any Variety or Volume focus always leads to the question of how to conduct Big Data Analytics. Velocity however is less about technology and more about the possibilities, about performance and about business impact. Being able to rapidly process and analyze vast amounts of unstructured data is crucial.

Consumers and citizens expect immediate response. They tweet their messages about what's on their minds and want to be adequately served by webcare teams. The conclusion is obvious: as long as it can be faster, new Big Data technologies and applications continue to evolve. Time is money and acceleration is synonymous with competitive advantage. Real-time isn't fast enough: predicting what will happen next is the real ambition, moving from predictive to prescriptive analysis.

2 Big Data transformation

The ideal data scientist is a much discussed colleague. This all-rounder must embody all the competencies for an organization to become Big Data driven, a transformation that requires direction and support from the top. For instance to start a statistical agency using mobile phone data. Insights from location data and people moving around have proven to be an excellent alternative for charting consumer behavior. But every time Big Data comes in there will remain much resistance and conservatism to counter, since as Clay Shirky puts it:

> *"Institutions will preserve the problem to which they are the solution."*

BI experts for instance, may have many solutions for problems that they cherish: traditional Business Intelligence and Big Data Analytics still are worlds apart.

3 Own your data and your future

It could be a time bomb under many Big Data initiatives: consumers and citizens who rule their own data. This scenario was discussed by an international group of CIOs at the Sogeti Big Data Summit 2012 where Doc Searls, author of *The Intention Economy: when customers take charge*, hosted the meeting. Searls argues that individuals themselves are in the best position to monetize their personal data, which might overthrow the relationship with organizations. Ultimately, Vendor Relationship Management would replace Customer Relationship Management. In his book *Who Owns the Future?* Jaron Lanier also pleads that people should make money from their own personal data. For Lanier this could be a solution to the Big Data Analytics problem that people will lose their jobs in a world af advanced smart systems and devices.

4 Privacy in the picture

How do we deal with privacy or what is left of it? When The Guardian started publishing Edward Snowden's revelations about the Big Data practices of secret services, it was fuel to an existing fire. The privacy issue always has been top of mind when personal data were involved. We now follow the great privacy debate on a daily basis in the media as we are painfully aware of the fact that no secrets are safe since agencies and organizations have access to much more information than we would like.

Organizations and governments of course have been collecting enormous amounts of data that can be related to an individual, but this kind of Personally Identifiable Information (PII) is protected throughout the world by privacy laws. However, in the digital age, legislation alone is insufficient by far and thanks to Big Data Analytics non-personal data also easily can lead to the right prospects. Few organizations seem to master the *Privacy by Design* maxim but the best advice to be trusted is: be transparent, comply and explain as much as you can.

5 Welcome to the No More Secrets era

The question remains how much the Big Data future is influenced by the Snowden revelations which have put data protection and privacy at the center of our attention. All major U.S. online services were persuaded to participate in the largest ever monitoring of data traffic: Big Data Analytics to the max. Friendly powers and foreign companies were tapped as smartphones and tablets were shamelessly searched. Encryption and other security systems were cracked or had loopholes to circumvent them. A fundamental undermining of computer, data and network security, plus of privacy and data protection. No one was aware of the extent and depth of these operations.

There are no secrets. In January 2010, Facebook CEO Mark Zuckerberg stated that the privacy era was completely over. In March 2012, CIA chief David Petraeus conceded that the relationship between identity and secrecy should be fundamentally discussed as all secret services have the task of being "world-class Big Data swimmers" in order to counter terror activities:

> *"Transformational is an overused word, but I do believe it properly applies to these technologies [. . .] Taken together , these [Big Data] developments change our notions of secrecy and create innumerable challenges – as well as opportunities."*

Do you copy? Challenges first! The whole security world was in distress and European politicians threatened with the suspension of international treaties. Via malware in Belgacom's network systems the British even appeared to have spied out NATO, while their supervising National Security Council was never informed.

Intelligence agencies are above the law, so much was clear, but that of course should have hardly come as a surprise. Still, no reasonable person can explain or justify this digital game of states within states. In September 2013, a speech of Sakharov Prize finalist Edward Snowden was read before the European parliament by National Security and Human Rights Director Jesselyn Radack, while the demand for well-shielded European cloud services resonated more than ever.

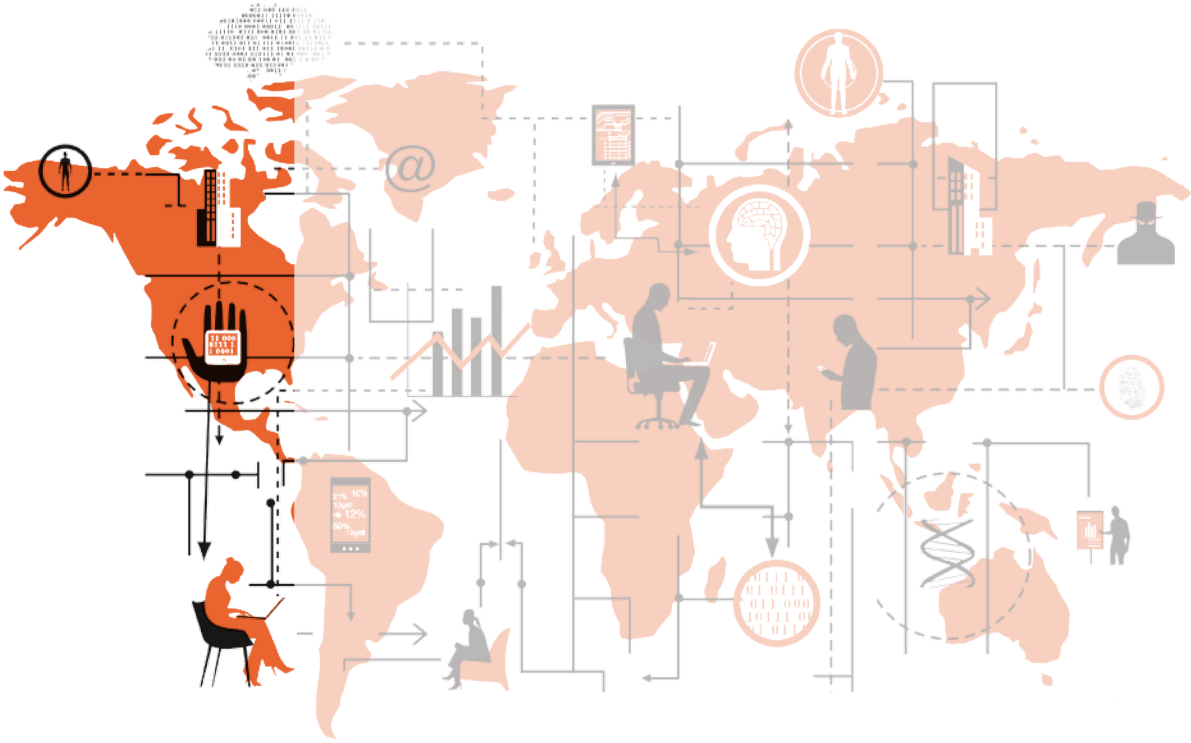
Forrester Research estimated the cost of the Snowden disclosures at $180 billion, in particular for the U.S. cloud computing industry. In October, all major online U.S. players signed a letter aiming to diminish the power of the U.S. National Security Agency NSA that had acted as if there were *No Such Agency*. Relationships between countries and between business, media and governments will have to be restored.

If privacy, or what's left of it, can be sustained in the explosion of Big Data capabilities remains to bee seen. But they also achieve breakthroughs in science, business economics and customer satisfaction. To quote the British science fiction author, inventor and futurist Arthur C. Clarke: *"Any Sufficiently advanced technology is indistinguishable from magic."*

We hope that *No More Secrets* let you walk firmly with both feet on the ground in the reality of Big Data Analytics, and that this book will continue to inspire you to create your own "magic moments" in search of better insights and business decisions.

Part I

Creating Clarity with Big Data



1 Digital data as the new industrial revolution

In 2012, approximately forty years after the beginning of the information era, all eyes are now on its basis: digital data. This may not seem very exciting, but the influx of various data types, plus the speed with which the trend will continue, probably into infinity, is certainly striking. Data, data and more data: we are at the center of an expanding data universe, full of undiscovered connections. This is not abstract and general, but rather specific and concrete, as each new insight may be the entrance to a gold mine. This data explosion is so simple and fundamental that Joe Hellerstein of Berkeley University speaks of 'a new industrial revolution': a revolution on the basis of digital data that form the engine of completely new business-operational and societal opportunities.

At the beginning of May 2012, at the Cloud Computing Conference arranged by Goldman Sachs, Shaun Connolly from Hortonworks presented data as "The New Competitive Advantage." Connolly articulated seven reasons for this statement, two of which were business-oriented, three were technological, and two were financial:

Business reasons

1. New innovative business models become possible
2. New insights arise that give competitive advantages

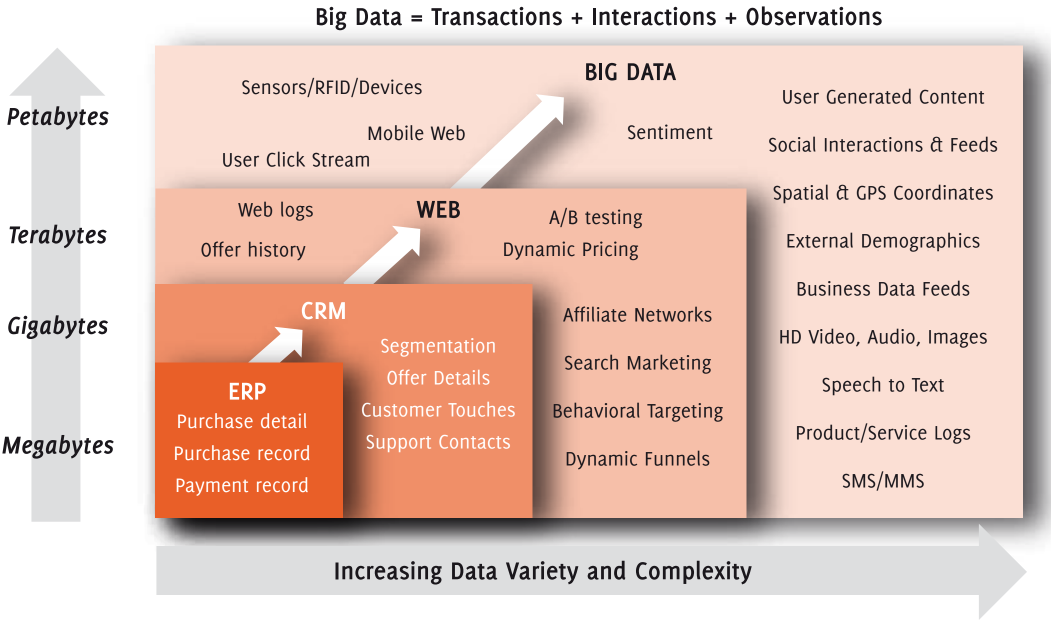
Technological reasons

3. The generation and storage of data continue to grow exponentially
4. We find data in various forms everywhere
5. Traditional solutions do not meet new demands regarding complexity

Financial reasons

6. The costs of data systems continue to rise as a percentage of the IT budget
7. New standard hardware and open-source software offer cost benefits

Connolly believes that, as a consequence of this combination, the traditional data world of business transactions is now beginning to merge with that of interactions and observations. Applying the formula *Big Data = Transactions + Interactions + Observations*, the goal is now: more business, higher productivity and new commercial opportunities.



Source: Contents of above graphic created in partnership with Teradata, Inc.

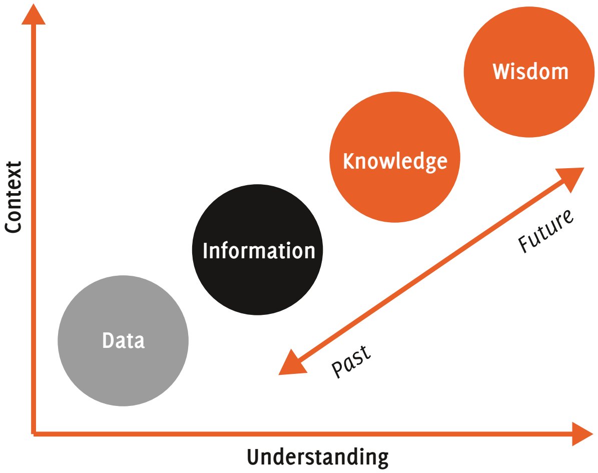
Digital data as the basis

At present we are living in at least three periods that build upon digital data: the information era, the social era, and the Big Data era. This is what is stated in Wikipedia's *List of Periods*, which covers the entire history of humankind. The explosive growth of data genuinely comes from all corners: from business transactions, mobile devices, sensors, social and traditional media, HD video, cloud computing, stock-and-share markets, Web-clicks, et cetera. All data is generated in the interaction between people, machines, applications and combinations of these. Those who have difficulty in grasping all this should take a look at a publicly accessible corner of our new data universe: the Linked Open Data domain at http://lod-cloud.net. The visualization of that data network and its components immediately clarifies what is going on in the world, in all sectors of the economy, society and science, and also in a combination of these.

Everything is information

Organizations exist thanks to information, and within the realm of science nowadays there is a movement that claims that, in fact, everything is information. Data forms the fundament of this information, and the more relevant facts we have, the better we can understand the most diverse issues, and the better we can anticipate the future. This is necessary in order to be able to take the correct decisions, certainly in these times of hypercompetition and crisis. The unprecedented data intensity of the Big

Data age that we have just entered, ironically at this crisis-ridden moment, is nevertheless a blessing, say the proponents. After all, analysis of complete datasets is, by definition, the only real way to be able to fully comprehend and predict any situation. This argument has no leaks, and thanks to modern and affordable IT – hardware, software, networks, algorithms and applications – analysis of complete datasets can now genuinely take off.



Big Data case: loss of clients

Until recently we were compelled to take random samples and analyze them. But how do you sample a network or a collection of subnetworks? If a telecom provider wishes to have insight into the circumstances under which a subnetwork of friends and acquaintances suddenly switches to a rival company (it "churns"), we are probably dealing with a total of more than 10 million existing and recent subscribers, with information on their habits, their expenditures on services, and who their friends are: the number of times the phone is used for calls or SMS messages, for example. We are dealing with tipping points: a part of the subnetwork churns and the rest follow after a (short) time. In itself, this is rather predictable: if colleagues or friends have switched and are better off or cheaper out under a rival, then there is a social and economic stimulus to switch as well. A provider will, of course, attempt to prevent this situation arising and must take a hard look at all the data. For example, if a random sample is taken from a million clients, the circles of friends that formed the basis of the switch can no longer be seen as a unit, and therefore in this case the basis for accurate prediction crumbles. Therefore, sampling is not the appropriate method here. In order to obtain a good view of the tipping points we must examine all the data in their proper context and coherence. Then, on the basis of developing patterns, we can anticipate their churn at an early stage and apply retention actions and programs.

Detection of fraud

Another area for which we require a complete dataset is fraud detection. The signal is so small that it is impossible to work with random samples until the signal has been identified. Accordingly, all data must be analyzed in this field as well. It can justifiably be referred to as an evident case of Big Data when the possibility of 'collusion' is being examined: illegal collaboration that is directed toward impeding others as much as possible and of sabotaging them, as occurs in the casino world. Churn and fraud detection are examples of the application possibilities of Big Data Analytics (see also Section 7).

Big Data Success Stories

In October 2011, under the title *Big Data Success Stories*, IBM published an illustrative reader with twelve different case studies, to demonstrate what Big Data actually signifies. We shall also respond to that issue here, in the following section and in Section 7, "Big Data in organizations in the year 2012." For the moment we shall proceed from the fact that Big Data Analysis goes further than what traditional relational databases can offer, and that current trends are moving toward the use of an increasing number of new data types. With all the empirical data that are there for the taking, it seems as if, in the future, we will only need to examine the facts in a smart way so that theory and model-forming, as intermediate steps, can ultimately be skipped. This Big Data promise was articulated as far back as 2008, in an article entitled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete".

2 Total data management in each organization

Big Data, the enormous data expansion that is spreading rapidly in all respects, demands total data management in every organization. This fact has been underlined by many experts, including The 451 Group.

An increasing quantity of data is arriving from all kinds of sources: from traditional transactional data to sensors and RFID tags, not forgetting social media, Internet, clouds and mobile devices. It no longer matters whether data is structured, semi-structured or unstructured, as current IT solutions for data acquisition and processing, and their affordability are thriving at the same time.

Data growth surpasses Moore's Law

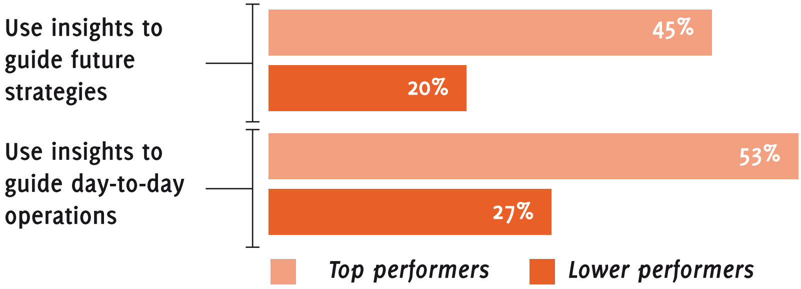
Although the flood of data now exceeds Moore's Law – every 18 months a doubling of the number of transistors per square inch on integrated circuits takes place against increasingly lower cost – we are still able to deal with the data deluge in a meaningful way. This is possible due to advanced hardware, software, networks and data technologies. In short, we are capable, along with a well-trained workforce, of exploiting the entire data field. Anyone who can do this well, stated Gartner in their presentation entitled "Information Management Goes 'Extreme': The Biggest Challenges for 21st Century CIOs", can achieve 20% better than competitors who do not do so:

> *Through 2015, organizations integrating high-value, diverse new information types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20%.*

The rules of the game remain the same, but the tactics have changed. Just as always, we wish to process information from raw data and extract intelligent new insights that enable better and faster business decisions. Big Data is actually an appeal to organizations to elevate their Business Intelligence efforts to a radically higher level: on the basis of the appropriate technology, the proper processes, the right roles and the relevant knowledge and expertise, called 'Data Science'. This must be practised throughout the entire organization and constantly.

Big Data is a new phase

Big or Total Data constitutes a new phase in the trend that was quantified by MIT Sloan Management Review and the IBM Institute for Business Value in 2010, in their study entitled *Analytics: The New Path to Value*. Almost half of the best-achieving organizations, it turned out, used digital data for their long-term strategies, in contrast to only a fifth of the underperformers. With regard to daily operations, this was even more than half of the top-performers against slightly more than a quarter of the poorly achieving organizations. The conclusion must be drawn that priority must be given to an analysis of the full array of digital data.



Of course, organizations do not wish to discard this type of advice, all the more because it builds logically upon existing Business Intelligence and the target of economic profit. But various demands and requirements must be dealt with and put in place. In addition to the potential and promise of Big Data, we shall also cover this aspect. The ambition of this book is to exchange thoughts with you on the topic of this important subject matter, and to jointly explore the possibilities.

3 Participate in our Big Data discussion at vint.sogeti.com

The Big Data issues about which we would like to exchange ideas and experiences, on the basis of this book, are threefold:

A. Your Big Data profile: what does that look like?
B. Ten Big Data management challenges: what are your issues?
C. Five requirements for your Big Data project: are you ready?

Note:

Interaction on this and related matters occurs on our website, and also face-to-face as far as we are concerned. We shall share new research insights with you on a weekly basis, via Blog posts, e-mail and Twitter alerts. The accompanying video material, presenting leading experts, is intended as inspiration to think through and discuss the entire theme of Big Data from various angles.

A. Your Big Data profile: what does that look like?

Big Data is concerned with exceptionally large, often widespread bundles of semi-structured or unstructured data. In addition, they are often incomplete and not readily accessible. “Exceptionally large” means the following, measured against the extreme boundaries of current standard IT and relational databases: petabytes of data or more, millions of people or more, billions of records or more, and a complex combination of all these. With fewer data and greater complexity, you will encounter a serious Big Data challenge, certainly if your tools, knowledge and expertise are not fully up to date. Moreover, if this is the case, you are not prepared for future data developments. Semi-structured or unstructured means that the connections between data elements are not clear, and probabilities will have to be determined.

B. Ten Big Data management challenges: what are your issues?

1. How are you coping with the growing quantities of semi-structured and unstructured data? It has been estimated that 80 per cent of the data most valuable to organizations are located outside the traditional relational data-warehousing and data-mining to which Business Intelligence has been primarily oriented until now.
2. Those new valuable data come from a range of different data types and data sources. Do you know which of these are important for your business and do you have a plan to apply them strategically?
3. Do you have an overall view of the complexity of your data, either independently or in combination? And do you know what exactly you want to know in which order of sequence. Now and in the future?
4. New insights obtained from the combination of structured and unstructured data may have an imminent expiry date. Are you aware of the desired speed of processing and analyzing various data and data combinations? Which issues that you might wish to solve require a real-time approach? Please keep in mind that real-time processes are needed to enable real-time decisions.
5. Have you thought about the costs of your new data management? How are they structured: according to data domains, technology and expertise, for instance?
6. The storage of all data that you wish to analyze and stockpile will probably make new demands upon your IT infrastructure. Do you have any kind of plan to deal with this, and are you also watching performance?
7. What is the state of your data security system?
8. The storage and security of Big Data is of major importance with regard to your data governance, risk management and compliance. Are you involving the appropriate departments and people in your Big Data activities?
9. Generating new business insights from large quantities of data requires an organization-wide approach. New knowledge and expertise are needed for this. Are they available in your organization and how can these be guaranteed and further developed?
10. Do you know what your Big or Total Data efforts mean for your energy use?

C. Five requirements for your Big Data project: are you ready?

On the basis of the above-listed management challenges, we now summarize five fundamental conditions that are collectively needed in order for you to embark confidently on a concrete Big Data project:

1. Your organization has at its disposal the right mindset and culture. There is no doubt, throughout the whole organization, about the usefulness of a Big or Total Data approach, you know where you want to begin, and what the targets for the future are.
2. There is sufficient management support and it is evident who the executive sponsors are.
3. The required expertise and experience with regard to Data Science and Big Data frameworks and tools are available and guaranteed.
4. Sufficient budget has been allocated for the necessary training, in order to ensure that the required expertise and experience, mindset and culture will bond.
5. There are adequate resources and budget for the development of Big Data applications, and you have selected the right partners and suppliers in this context.

4 Why the word "Big"?

We refer to something as "big" – Big Mac for example – to draw attention to its volume. But if we supply no relevant image, the word "big" immediately evokes fundamental questions. That is also exactly the case with Big Data, and also with the related Big Science. How large is Big Data actually, and in relation to what?

"Big" is not a particularly handy term

Accordingly, the analysts at Forrester and at Gartner agree completely with this statement: in retrospect, "big" is perhaps not a convenient name for the flood of data that is increasing at an enormous pace. Both research companies, and others with them, prefer to use "extreme" rather than "big." That term also has a longer history in the field of statistics.

In everyday life, “big” refers to very concrete oversize phenomena. But inconceivably high quantities of digital data are not perceived by the eye. In addition, more is happening than “quantity” alone.

Big Data and Web 2.0

It is no coincidence that O’Reilly Media introduced the term “Big Data” a year after Web 2.0 appeared, as many valuable Big Data situations are indeed related to consumer conduct. Web 2.0 provided the impulse to rethink the interaction that was taking place on Internet, and to push it somewhat further. In much the same way, the qualification “Big Data” calls attention to the business possibilities of the flood of data on the one hand, and the new technologies, techniques and methods that are directed toward these, on the other.

A simple answer

As mentioned, the increase in data has now exceeded Moore’s Law. Various types of data in combination with the necessary speed of analysis now form the greatest challenge, although we must not forget the limited number of people who can deal proficiently with Big Data. In 2020, there will be 35 zettabytes of digital data. That represents a stack of DVDs that would reach half way from the Earth to Mars. Facebook has 70 petabytes and 2700 multiprocessor nodes. The Bing search engine has 150 petabytes and 40,000 nodes. But what does Big Data exactly signify for organizations? We can approach Big Data from the standpoint of the issues, but also from the standpoint of the solutions. The simplest response comes from Forrester Research and is as follows:

> *Big Data: Techniques and Technologies that Make Handling Data at Extreme Scale Economical.*

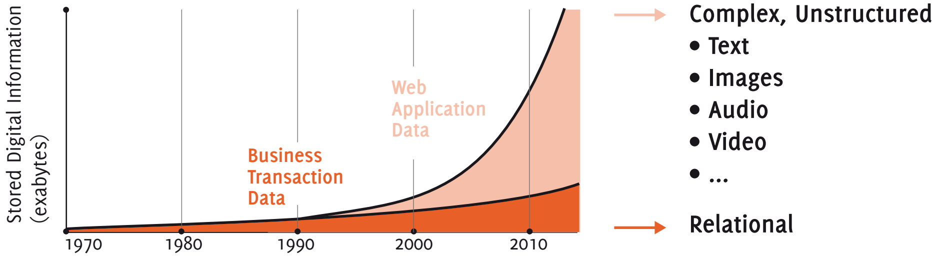
Just like The 451 Group and Gartner, Forrester also makes no distinction between Big and Little Data. Compared to bygone times, many new and different data have arrived on the scene, and this is an ongoing process; but data remain data. They go hand in hand, and we can only truly advance further if there is well-thought-out integration of the whole spectrum of various orders of magnitude. We are dealing with a single data spectrum, a single continuum, and that is what organizations ought to be strategically exploring step by step.

One large data continuum

Around thirty years ago, this also applied to the growth of scientific activity: large and small. In his book entitled *Reflections on Big Science* (1967), the atomic scientist Alvin Weinberg wrote:

> *The scientific enterprise, both Little Science and Big Science, has grown explosively and has become very much more complicated.*

This observation referred to science at that time, and it now refers precisely to what is happening in the realm of data. Check what Chirag Metha has to say. As a Technology, Design & Innovation Strategist, Metha was associated with the Office of the CEO at SAP:



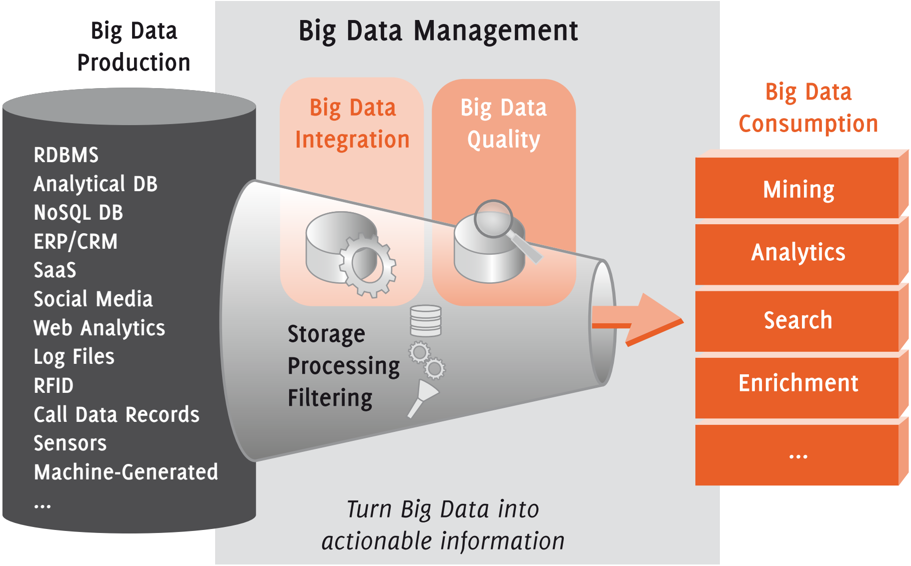
Big Data does not at all mean to say that we ought to forget Little or Small Data, or Medium, Large et cetera. On the contrary, it is important that we can and must review all the data in all their forms. It is possible technologically, and desirable, if not essential, businesswise.

> *Today, technology — commodity hardware and sophisticated software to leverage this hardware — changes the way people think about small and large data.* ***It's a data continuum.*** *[...] Big Data is an amalgamation of a few trends – data growth of a magnitude or two, external data more valuable than internal data, and shift in computing business models. [...]* ***Big Data is about redefining what data actually means to you.*** *[...] This is not about technology. This is about a completely new way of doing business where data finally gets the driver's seat.*

This is particularly the case because 80 per cent of all new data is not relational or is unstructured and, in combination with transaction data, contains the most valuable information for organizations. In the view of some people, not all data that initially seem unstructured need to remain so, not by a long way, and indeed such data can be accommodated within a structure with relatively little difficulty.

5 The importance of Big Data

The reason why we should wish to have and examine all that data is evident. Social media, web analytics, logfiles, sensors, and suchlike all provide valuable information, while the cost of IT solutions continues to drop and computer-processing power is increasing. With developments like these, the surplus of information seems to have largely vanished: in principle, organizations are now capable of managing the flood of data and to use it to their own (financial) advantage. Those who excel in acquiring, processing, and managing valuable data, says Gartner, will be able to realize a 20% better result, in financial terms, than their competitors.



Within organizations, the share of unstructured data, such as documents, e-mail and images, is around 60 to 80 per cent. Of all data analyses that currently take place in organizations, 5 to 15 per cent contain a social component that enriches the structured data. This number must increase, not least because of all the external data that can be included in the analyses.

The Internet of Things is also becoming an increasingly rich source of data. At this moment, says Cisco CTO Padmasree Warrior, there are 13 billion devices connected to the Internet and that will be 50 billion in 2020. IDC expects more than 1 billion sensors to be connected to the Internet by that time. All the accompanying data flows can supply interesting insights that can aid better business decisions.

We are at Big Data's point of departure

Banks belong to the top of the organizations that are engaged with Big Data but, in the report with the eloquent title *Big Data: Harnessing a Game-changing Asset* by the Economist Intelligence Unit, Paul Scholten, COO Retail & Private Banking at ABN AMRO, candidly admits that the bank is in an exploratory phase when it comes to making good use of unstructured social data in particular:

> *We are used to structured, financial data. […] We are not so good at the unstructured stuff. […] The company is just beginning to understand the uses of social media, and what might be possible in terms of improving customer service.*

Mark Thiele states that it is interesting to compare Big Data in the year 2012 with the start of the World Wide Web. Thiele is the Executive VP Data Center Technology at Switch, the operator of the SuperNAP data center in Las Vegas, the largest and most powerful of its type in the world:

> *Big Data today is what the Web was in 1993. We knew the Web was something and that it might get Big, but few of us really understood what "Big" meant. Today, we aren't even scratching the surface of the Big Bata opportunity.*

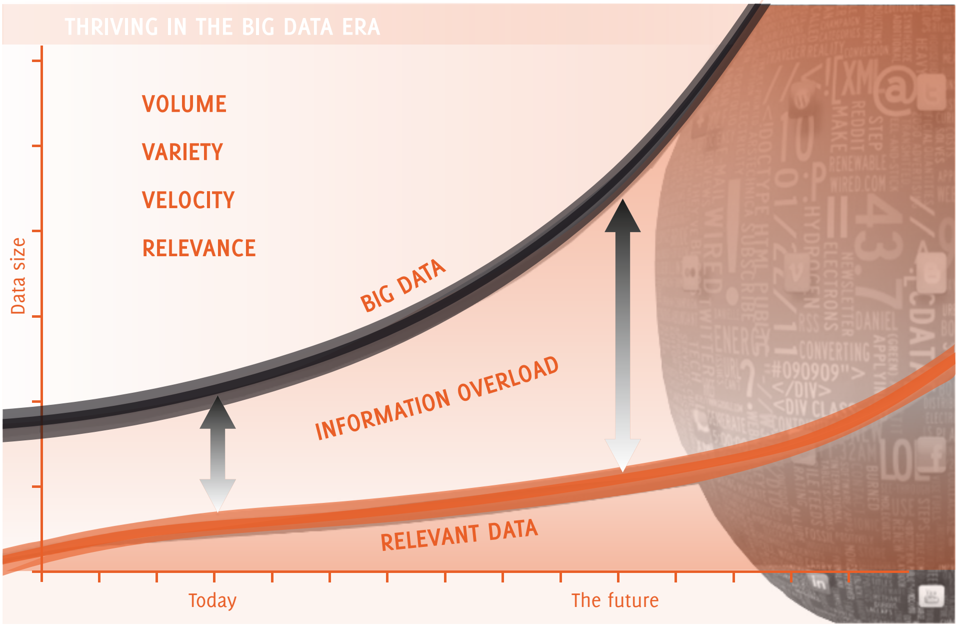
No isolated phenomenon

If there is one thing that has become clear, that is the fact that Big Data is not an isolated phenomenon. The word "big" emphasizes the quantitative aspect which fortunately immediately raises the necessary questions, so that we are compelled to think more profoundly about Big Data.

In March 2012, Credit Suisse Equity Research published the report entitled *The Apps Revolution Manifesto, Volume 1: The Technologies*. The authors regard, in particular, the convergence of Service-Oriented Architecture, Cloud, Fast Data, Big Data, Social and Mobile as being determinative of the value that new enterprise applications can provide. Credit Suisse Equity Research estimates this development to be just as transformative as the client/server and web applications were in the past.

Volume, Variety, Velocity

As far back as 2001, Doug Laney made clear – then at META Group and nowadays at Gartner – that three factors can influence one another in the growth of data flow: the quantity of data *(Volume)*, the nature of the data type: structured, semi-structured and unstructured *(Variety)* and the desired analysis rate *(Velocity)*. Nowadays we often add *Complexity*, *Value* and *Relevance* to this list. The last two are included because we would like to know what we can and want to do with all the data, so that we are not investing time, money and effort for no return.

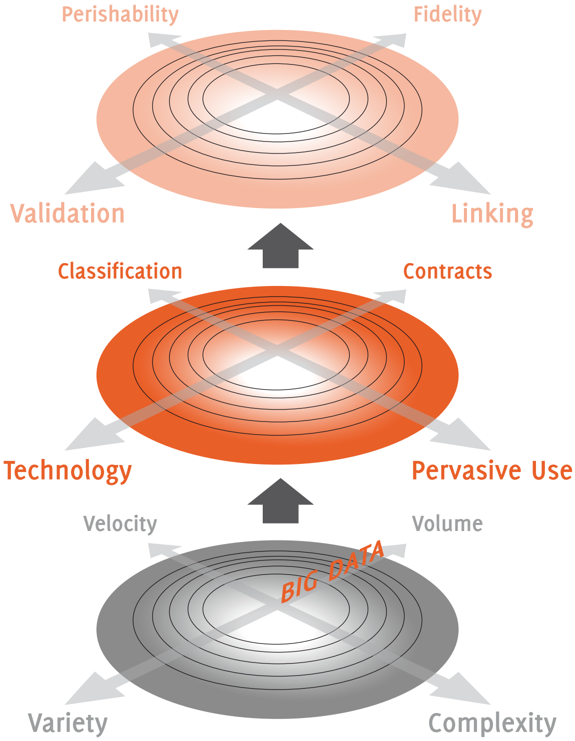


Big Data as the next frontier

On that basis, predicts the McKinsey Global Institute in its report entitled *Big Data: The Next Frontier for Innovation, Competition and Productivity*, the right exploitation of Big Data can produce hundreds of billions of dollars for various sectors of the American economy. McKinsey underlines the great sectoral differences (see Section 11) with respect to the ease with which Big Data can be acquired, set against the value that the use of Big Data can be expected to produce. It further emphasizes the necessity of eradicating the knowledge gap in organizations, with regard to dealing with (Big) Data (see Section 10).

6 Big Data is Extreme Information Management

Gartner has now elaborated the basic model of Volume, Variety and Velocity into three interactive layers, each with four dimensions (as shown in the illustration). The resulting twelve factors dovetail together and must all be purposefully addressed in the information management of the 21 st century: separately and as a whole.



In short, here we have, moving from the bottom to the top, the following: departing from the variety and complexity, in particular, of an increasing amount of data – often also in real-time – it is very possible to express validated statements and to establish connections on the basis of correct technological applications in combination with intensive input of all data, in order to elevate business decision making to a qualitatively higher level.

If we take Big Data as the point of departure, we find ourselves on the volume side, as the name indicates. Variety and speed are the other dimensions at that level. An extra addition is the complexity of not only the data but also of the 'use cases': the way in which all data is brought into association by means of relevant and constructive questioning. We have already presented a concrete typology on the basis of the formula *Big Data* = *Transactions* + *Interactions* + *Observations* in Section 1.

The intermediate level is concerned with access and control. To start with, there are always agreements (*Contracts*) about which information precisely (*Classifica-*

tion) should be recorded and how it can be used. Social media and cloud computing provide splendid opportunities, but new technology (*Technology*) is needed to ensure that the data can be used everywhere and at any time (*Pervasive use*).

The top layer covers the reliability of information (*Validation, Fidelity*). It must be not only relevant and accurate when acquired (*Perishability*), but also in the use case. It is also important whether or not enrichment occurs in combination with other information (*Linking*).

Altogether, in a Big Data context, organizations must respond to the six well-known standard questions: what, when, why, where, who and how? The first four cover the structure of your Enterprise Information Architecture and the last two that of your Enterprise Information Management.

What? What are the correct data and information?
When? What are their ideal lifecycle stages?
Why? What are the right characteristics?
Where? What are the proper interfaces for interaction?
Who? What are the right roles in the organization?
How? What are the right information activities?

This is the concretization that belongs to the standard questions, in a nutshell. These questions serve as a guideline for the further structuring of Big Data, Total Data or Extreme Information Management processes.

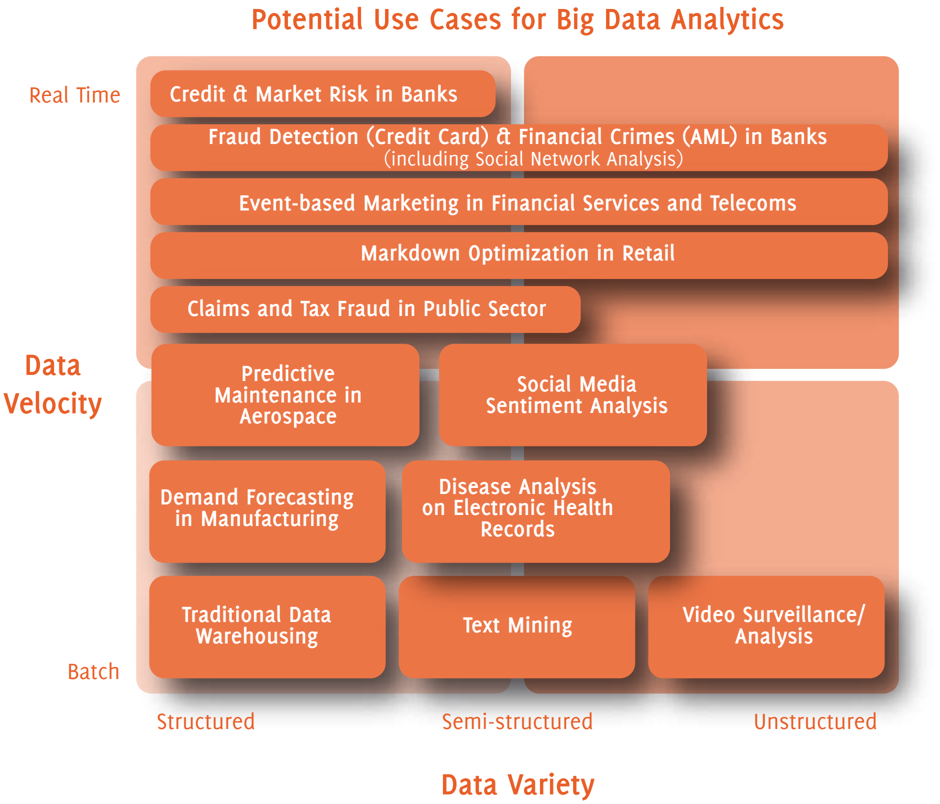
EIM and Big Data governance

IBM's Big Data Governance Maturity Framework provides reliable handholds for Extreme Information Management. The accompanying checklist contains more than 90 points of interest in 11 subareas. This elucidating material can be accessed via:

ibmdatamag.com/2012/04/big-data-governance-a-framework-to-assess-maturity

7 Big Data in organizations

Along the axes of speed (Velocity) and data types (Variety) – thus deliberately abstracting from data quantities (Volume) – SAS and IDC formulated the following self-evident potential of Big Data Analytics for organizations.



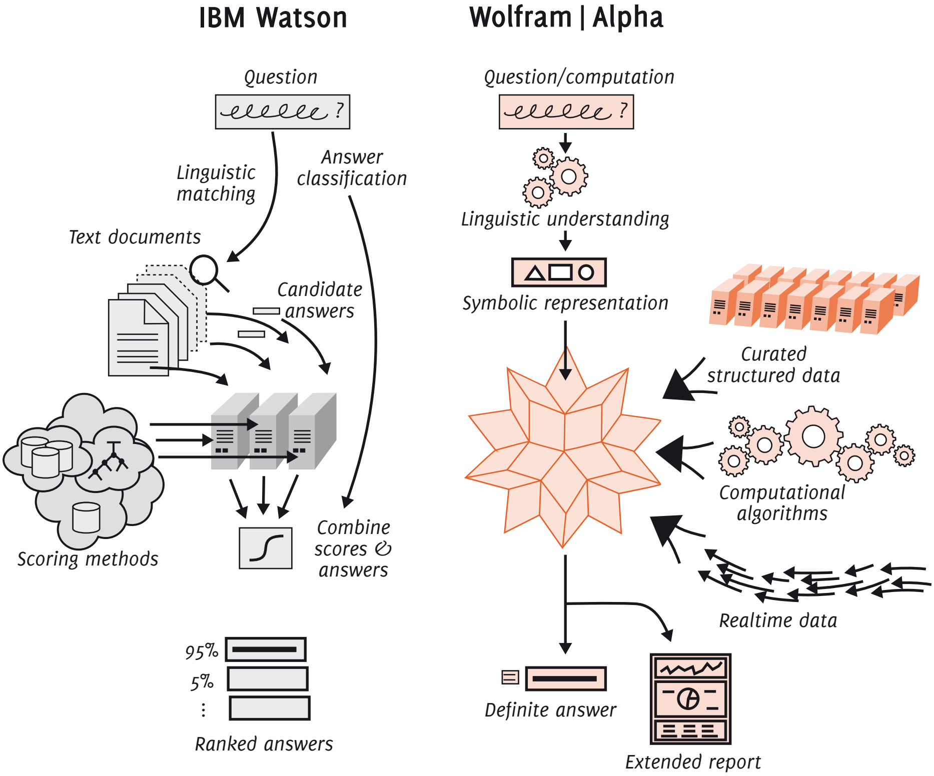
Potential Use Cases for Big Data Analytics
Real Time
Credit & Market Risk in Banks
Fraud Detection (Credit Card) & Financial Crimes (AML) in Banks (including Social Network Analysis)
Event-based Marketing in Financial Services and Telecoms
Markdown Optimization in Retail
Claims and Tax Fraud in Public Sector
Data Velocity
Predictive Maintenance in Aerospace
Social Media Sentiment Analysis
Demand Forecasting in Manufacturing
Disease Analysis on Electronic Health Records
Traditional Data Warehousing
Text Mining
Video Surveillance/ Analysis
Batch
Structured
Semi-structured
Unstructured
Data Variety

Data Science as a sport

The desired intensive interplay between staff members in the field of Big Data and the current shortage of expertise and experience within organizations allow scope for the Web 2.0 approach called ‘crowdsourcing’. The Australian Kaggle.com is one example of this kind of online initiative in Big Data service-provision. It makes a sport of Big Data challenges: “We’re making data science a sport.” In their online arena, as Kaggle calls it, data cracks can participate in diverse competitions. Organizations offer their data and questions, which are subsequently and skillfully analyzed right down to the finest details by experts affiliated with Kaggle. The best solution is the winner and is awarded the stated prize. Fame, prize money and game enjoyment are what the gladiators are seeking:

> *Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.*

Developments such as Kaggle are very interesting because the potential of innovations and/or innovative entrepreneurship on the basis of Big Data are highly valued. State-of-the-art computer systems such as Watson by IBM and Wolfram|Alpha play a major role here. These and other intelligent computers are applied in an increasing number of Big Data challenges: from banks to the Smart Grid and healthcare.



The Social Business Analytics example of churning, the erosion of a client stock, which occurs all too frequently in for instance the telecoms industry, was dealt with at the beginning of this part, in Section 1.

The Smart Grid

All over the world, a great number of pilot projects are currently taking place at the interface of Big Data and the so-called 'Smart Grid'. Grid monitoring is one of the major areas of interest, as is now happening in the Tennessee Valley Authority project, in which 9 million households and more than 4 billion measurements a day collectively supply 500 terabytes of data. Typical applications include the tracing of interruptions and the monitoring of energy use. There are smart meters for electricity, gas and water. It is expected that 270 million will be operational in 2013. If we take this a step further, to intelligent houses, these will each generate 4 to 20 petabytes of data a year on the basis of 100 sensors per household. The need for Big Data applications in the utilities sector is thus increasing, and evolving deregulation is fueling this trend.

Healthcare

Healthcare is a broad domain that affects us all directly. With regard to clinical use of Big Data, thus for healthcare treatment, it is beneficial to be able to follow information that has been compiled in all sorts of ways over the course of time. In addition, a beginning can be made on pattern recognition, particularly the detection of events that do not occur frequently or are not perceptible when research is oriented to small populations. A good example is the way in which Google is capable, by means of Big Data analysis and in real-time, of following the way a flu epidemic is spreading. Even more impressive is the way in which the scientific Global Viral Forecasting project uses Big Data to prevent worldwide pandemics such as HIV and H1N1 infection. In such matters we must be aggressively proactive, as the absence of results has taught us that we simply cannot just sit and wait while potential catastrophes are developing all around us.

Ahead of our gene chart

A fundamental Big Data development in the field of healthcare is the ambition of the Broad Institute, an initiative of MIT and Harvard, to expand the Human Genome Project, which was eventually rounded off in 2003. Over a period of 13 years, scientists ultimately managed to chart all the 20,000 to 25,000 genes plus the 3 million basic pairs of human DNA. What the mega-project primarily proved was that genes only make up a minor part of our genome and that there are many more fundamental elements that must be identified and investigated.

The Broad Institute has been engaged with this assignment since 2003, and particularly with the issue of how cells actually process information, which not only leads to a better understanding of the genome but also has great therapeutic value. In combination with other institutes, the Broad Institute is currently researching the cell mutations that cause cancer, the molecular structure of the viruses, bacteria et cetera. that are responsible for infectious illnesses, and the possibilities of their use in the development of medicines.

Genome biology and the study of cell circuits belong to the most important Big Data challenges of our time. At the end of 2011, the Broad Institute had amassed 8 petabytes of data. The institute is continually working on dozens of specialist software tools in order to be able to analyze the data in the required way. All software and data can be downloaded by everyone.

Social Analytics

Warehouses use Social Analytics to rapidly adapt their online assortment to the customers' wishes on the basis of terabytes of search assignments, Blog posts and tweets. They now do so within a few days, instead of the six weeks that it normally used to take. Modern Social Analytics tools have been optimized for use by business professionals, and can cope with all kinds of data sources: publicly accessible sources, own data and that of partners.

The data flow revolution

Software for the analysis of data flows is used to uncover real-time deviations and new patterns in the data. In this way, organizations can immediately gain new insights and take quick decisions that are necessary on the basis of the latest developments. In this context, you can think of tweets that are monitored, or Blog posts, video images, electrocardiograms, GPS data, various types of sensors and financial markets. Modern data-flow software makes it possible to monitor real-time complex associations in situations that are much more complicated than relational databases and traditional analytical methods could possibly cope with. Ranging from patient care to better customer service, data-flow software offers surprising new possibilities.

Preventing medical complications

In hospitals, the respiration, blood pressure and the temperature of patients are continually monitored. In order to be able to detect the often subtle signals warning of complications, data-flow systems have to be applied. They are capable of identifying the first indicators of malfunction, well before the symptoms actually appear. In the past, 1000 measurements per second were aggregated to form patient reports every half hour or hour, but that is now considered as too crude. In this case, data-flow systems are of vital importance in order to be able to intervene proactively.

An optimum service

Another example is the service to customers. Internet and social media have empowered the customers and made them fastidious. On average, we trust one another's opinions three times more than we trust those expressed by corporate adverts. Therefore it is essential to listen attentively to what customers and others online have to say and to the information that they are exchanging. The improvement of service currently demands close attention to comments on websites, in e-mails, in text messages and on social media. If members of staff have to do that manually, the process is much too slow and there are too many inconsistencies in the reporting and the follow-up. With advanced data-flow software for content analysis, organizations are now capable of automatically analyzing that kind of unstructured data and of categorizing it according to certain terms and clauses that occur within the text. With such a policy, the car-hire company Hertz has doubled the productivity of its customer service.

Visionary phase

The examples given with regard to Big Data are as yet rather rudimentary. This is probably an indication of the phase we are in regarding Big Data. Organizations are not yet basing their distinctive value on their capacity to deal with Big Data. This far, we have not been able to identify the real "heroes" of this era, so that the disruptive potential only glimmers through the examples. We are currently in a visionary phase, in which much experimentation is going on. In this book VINT will pay particular attention to cases in different areas, from various angles and sectors.

8 With Big Data from Big Science to Big Business

Big Data is developing most rapidly in the world of Big Science. In 10 years, 2800 radio telescopes in the Square Kilometer Area project (SKA), the largest Big Science project ever, will generate 1 billion gigabytes of data daily. That is equal to the entire Internet on a weekday in 2012. As far back as 2008, Chris Anderson proclaimed the *Petabyte Age* in the magazine *Wired*, and Joseph Hellerstein, from UC Berkeley, announced the *Industrial Revolution of Data*. In comparison: in 2012, Google processes a total of 5 petabytes or 5000 terabytes per hour.

Big Data, Big Science and Big Bang

The terms Big Data, Big Science and Big Bang are all related to a completely different situation than the one to which we have traditionally been accustomed. For Big Bang, we can thank Fred Hoyle, the British astrophysicist, who coined the term in a radio broadcast in 1949. Atomic scientist Alvin Weinberg popularized Big Science in the *Science* magazine in 1961. And it was only relatively recently, in 2005, that Roger Magoulas of O'Reilly Media came up with the term Big Data. Its use was oriented to organizations: ranging from *Next Best Offer Analytics* directed toward the individual, to production environments and sensor data.

Big Business and Big Bucks

So, it is a good habit to call something "big" if we wish to draw attention to it. In this context we can think of *Big Brother* (1949) by George Orwell, not forgetting more profane matters such as Big Business – large (American) enterprises from the mid-nineteenth century – and Big Bucks, both of which have a direct association with Big Science and Big Data. With respect to Big Data, we are currently shifting from megabytes, gigabytes and terabytes to the vertiginous age of petabytes, exabytes and zettabytes. It's all happening extremely rapidly.

The notion that opportunities to capitalize on Big Data are simply lying there, ready to be seized, is echoing everywhere. In 2011, the McKinsey Global Institute called Big Data "the next frontier for innovation, competition, and productivity" and the Economist Intelligence Unit spoke unequivocally of "a game-changing asset." These are quotes taken from titles of two directive reports on Big Data, a topical theme that is developing vigorously, and about which the last word has certainly not been uttered. McKinsey states it very explicitly:

> *This research by no means represents the final word on big data; instead, we see it as a beginning. We fully anticipate that this is a story that will continue to evolve as technologies and techniques using big data develop and data, their uses, and their economic benefits grow (alongside associated challenges and risks).*

The Global Pulse project

As if he wished to underline the qualifying words of McKinsey, Ban Ki Moon, the Secretary-General of the United Nations, presented the so-called "Global Pulse project" at the end of 2011, geared to keeping up to date with a number of developments all over the world via large online datasets – *New Data* in Global Pulse terminology. The project is being run as a cooperative endeavor with various commercial and academic partners, with the ultimate aim of being able to intervene earlier and better in crisis situations if that should be necessary. There are five main projects:

1. A Global Snapshot of Well-being through Mobile Phones
2. Real-Time E-Pricing of Bread
3. Tracking the Food Crisis via Online News
4. Unemployment through the Lens of Social Media
5. Twitter and Perceptions of Crisis-Related Stress

Data Science rules!

Despite such indicative initiatives, the Big Data concept is most closely related to what we call Big Science. There, the Volume, Variety and Velocity aspects, in combination with state-of-the-art hardware and software, are most obviously present, although some people may contest scientific Relevance and Value, certainly in times of crisis. Moreover, the CERN particle accelerator and hypermodern telescopes are somewhat larger than what we have to deal with businesswise, and they are of a completely different order in terms of data techniques. So, how does Big Data bring us from *Big Science* to *Big Business*? The heart of the answer is *Data Science*, the art of transforming existing data to new insights by means of which an organization can or will take action.

Without mentioning the currently much-discussed concept of Data Science, Chirag Metha, the former Technology, Design & Innovation Strategist for the SAP Office of the CEO, emphasized above all the importance of the tools and the corresponding collaboration, as Big Data is certainly not only for experts. On the contrary, it is imperative to involve as many people as possible in the data chain:

> *Without self-service tools, most people will likely be cut off from the data chain even if they have access to data they want to analyze.* ***I cannot overemphasize how important the tools are in the Big Data value chain.*** *They make it an inclusive system where more people can participate in data discovery, exploration, and analysis.* ***Unusual insights rarely come from experts; they invariably come from people who were always fascinated by data but analyzing data was never part of their day-to-day job.*** *Big Data is about enabling these people to participate – all information accessible to all people.*

9 Big Data as new Data Science era

Right from the outset, a key characteristic of Big Science was the fact that the isolated scientist, working in his ivory tower, had become a thing of the past. But it did not remain a distinctive feature of Big Science, as co-operation soon became the norm across the whole of society. Modern science without well-coordinated collaboration has become inconceivable. The report entitled *Big Science > Big Data > Big Collaboration: Cancer Research in a Virtual Frontier*, dating from October 2011, emphasizes that from a Big Data perspective. In this book Big Science is put into the same category as Big Data and Big Collaboration. In the report itself, the three "Bigs" mentioned in the title are supplemented by *Big Technology* or *Big Compute*:

> *Big Science generates dimensions of data points and high-resolution images to be deciphered and decoded. In cancer research, Big Data often require on-demand Big Compute across settings using a private cloud, a public cloud or mix of the two.*

It is exactly this that changes for organizations when they decide to work with Big Data. If existing technologies and working methods in an organization are not able to cope with Big Data, a new approach will be needed. This means: investing in hardware, in people, in skills, in processes, in management and in governance. According to Gartner, Big Data is primarily literally the Volume component at the basis of what is referred to as *Extreme Information Management*. An integral part of that is *Data Science*, the "science" that inevitably enters the organization along with Big Data, Fast Data, Total Data and Dynamic Data. Chirag Metha gives the following profile sketch of a data scientist:

> ***The role of a data scientist is not to replace any existing BI people but to complement them.*** *You could expect the data scientists to have the following skills:*
>
> - *Deep understanding of data and data sources to explore and discover the patterns at which data is being generated.*
> - *Theoretical as well practical (tool) level understanding of advanced statistical algorithms and machine learning.*
> - *Strategically connected with the business at all the levels to understand broader as well deeper business challenges and being able to translate them into designing experiments with data.*

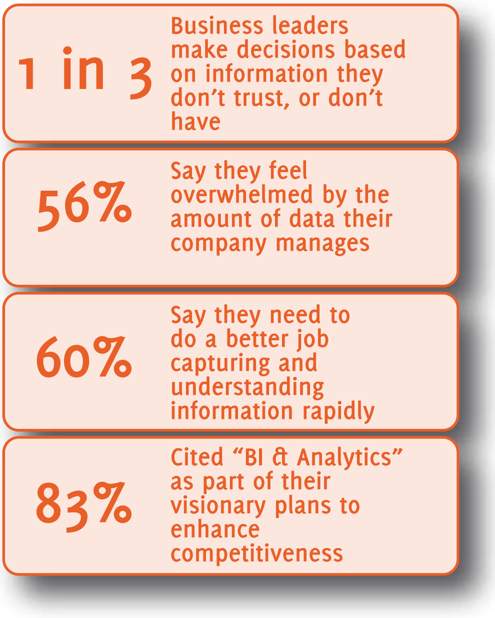
- *Design and instrument the environment and applications to generate and gather new data and establish an enterprise-wide data strategy since one of the promises of Big Data is to leave no data behind and not to have any silos.*

Big Data: a new microscope

With his *Principles of Scientific Management*, dating from more than a century ago, Frederick Taylor put the "scientization" of organizations on the agenda; in his particular case this was scientific management. This was important but it was essentially an issue of continuous improvement. With Big Data, the enthusiasts see a fundamental change, somewhat similar to the advent of the microscope. This is currently a favored analogy: we are on the brink of a new era, comparable with the beginning of modern science around 400 years ago. Owing to the digital "microscope", which is currently being invented for Big Data, as it were, we will soon be able to analyze and predict events much more scientifically and accurately in all fields, according to MIT professor Erik Brynjolfsson. Eventually we will be able to zoom in and out rapidly thanks to advanced hardware and software, with the ultimate aim of discovering structures and connections that enable us to obtain spectacularly better insight and solutions, and make better decisions: *Data Driven Decisions* and *Predictive Analysis.*

10 Closing the knowledge gap is essential

As a topical business theme, with sky-high economic and societal promise, Big Data is currently the subject of much interest and is gathering momentum. This will remain the case, at least in the near future, and accordingly there is a need for a clear picture. In that context, as the McKinsey Global Institute has calculated, 140,000 to 190,000 data experts (data scientists) will have to join organizations in the USA alone, and the number of business people who can deal with such data will have to increase by 1.5 million. First of all, a certain knowledge level is required in order be able to handle Big Data responsibly. Unfortunately there is a structural lack of knowledge in organizations across the entire spectrum. According to an IBM study dating from 2011, organizations are most willing to introduce structural improvements, as indicated by the percentages shown below. A few years ago, the excuse could still be applied that the development of Big Data was only possible for scientific people and a select number of organizations. For all other parties it was simply too difficult and too expensive. That is no longer the case. Pioneers such as Walmart, Tesco and Google have demonstrated that data can be the source of steady competitive advantage. According to IBM, no fewer than 83% of the CIOs currently nurture visionary plans to significantly improve the competitive position of their organization by means of new Business Intelligence & Analytics on the basis of Big Data.



1 in 3	Business leaders make decisions based on information they don’t trust, or don’t have
56%	Say they feel overwhelmed by the amount of data their company manages
60%	Say they need to do a better job capturing and understanding information rapidly
83%	Cited “BI & Analytics” as part of their visionary plans to enhance competitiveness

The Economist Intelligence Unit underlines this, but also subdivides Big Data conduct in large organizations into the following maturity quartet:

- **Data wasters**
  Of the data wasters, 30 per cent give no priority to the gathering of data. The 70 per cent from this category who do give priority to data-gathering use the data much too sparingly. Such organizations are below-average achievers. *We find them in every economic sector.*
- **Data collectors**
  These organizations recognize the importance of data, but do not have the resources to capitalize on them. They can only store them. They have immersed themselves completely in data. *We find this category primarily in healthcare and professional services.*
- **Aspiring data managers**
  This is the largest group. People are fully aware of the importance of Big Data for the future of the organization. They use data for strategic decision-making and make solid investments in that area. But they have never reached the upmost level

in terms of achievement. *We find them mainly in the communications branch and in retail services.*

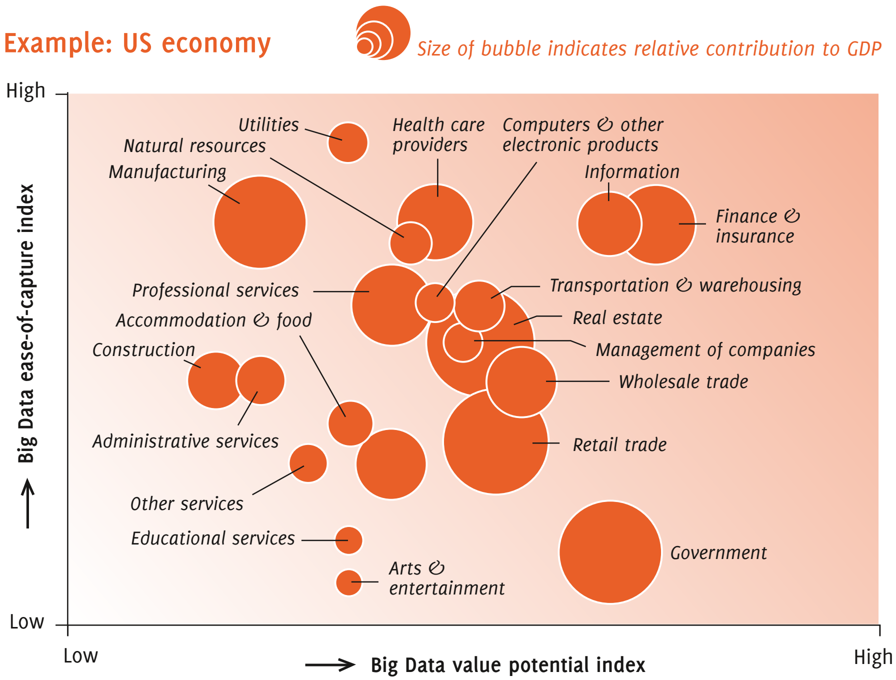
- **Strategic data managers**
  This is the most advanced group of Big Data users. These organizations first of all identify specific metrics and data that are related to their strategic targets. *We find them primarily in the manufacturing industry, in financial services and in the technology sector.*

Thus, organizations should not merely collect all kinds of data, but should also develop the wish and competence to work with as much data as possible. In conjunction with business professionals, data scientists must help interpret all the data and generate insights that are genuinely beneficial to the organization. This may concern specific issues or exploratory data research. The intention is to transform an organization from an intuitive decision-making instance into a data-intensive one, shifting from the *heroic manager* who takes decisions simply hoping for the best and knowing that there is too little data available, toward the more *scientific manager* who first seeks data and insight.

11 Big Data in hard cash

Precisely why Data Science skills are needed so badly has been quantified by McKinsey as follows. According to the office, trillions of dollars and Euros can be generated in value worldwide on the basis of Big Data. For example, 300 billion dollars in American healthcare, 250 billion euros in European government, more than 100 billion dollar in the American telecom business and up to 700 billion for their customers, can be earned on an annual basis. By capitalizing on Big Data, the American retail trade could increase net yield from turnover by more than 60 per cent, and the manufacturing industry would eventually only need to spend half on production development and assembly, while working capital could decline by 7 per cent.

These are examples from the overview picture of American economic sectors on the next page. The great sectoral differences between the ease with which Big Data can be obtained, set against the value that can be expected from using Big Data, are obvious. The McKinsey Center for Business Technology published this chart early 2012 in the reader *Perspectives on Digital Business*, on the basis of information from the report entitled *Big Data: The Next Frontier for Innovation, Competition, and Productivity* by the McKinsey Global Institute, May 2011.



Source: McKinsey Center for Business Technology (2012)

To determine the ease of obtaining data ("ease of capture") on the vertical axis, the researchers have investigated four factors: the analytic talent available, the IT intensity, the data-driven mindset, and the availability of data in a sector. The potential value (horizontal axis) is a function of the following five factors: the amount of data present, the variation in business-economic performance, contact with clients and suppliers, transaction intensity, and the competitive turbulence in a sector. The size of the circles in the figure indicates the relative contribution of a sector to the Gross Domestic Product.

Big Data has great potential particularly in areas that involve many people, such as utilities and healthcare. This is mostly so due to the relative ease with which Big Data can be obtained, as the figure above shows. In that context, utilities take the title. In terms of the combination of Big Data ease-of-capture, client relevance, financial profit and contribution to the economy, the information-processing industries, including financial service-providers, occupy top position.

12 Summary

Big Data is comparable to what the World Wide Web was in the early nineties. An enormous acceleration has taken place, everything is being connected to everything else, and the corresponding visions are being formulated. Many people expect that the current data focus will turn the world upside down: in terms of economics, society, innovation and social interaction.

Organizations are currently faced with the major challenge of having to imagine the concrete possibilities of Big Data. How could Big Data generate a revolution in your professional field? Or what would change if you truly succeeded in knowing everything you wanted to know? Could you cope with that? Would you like that and, if so, in which way? And can you allow yourself to wait for further developments in the realm of Big Data, or perhaps avoid participating altogether?

The core of Big Data is that we are dealing with one data spectrum, one continuum. Organizations will explore this continuum step by step, because we do not wish to ignore new possibilities to make better decisions. To help define the urgency of transformation within your organization, we presented and explained the following issues in Section 3:

A. Your Big Data profile: what does that look like?
B. Ten Big Data management challenges: what are your issues?
C. Five requirements for your Big Data project: are you ready?

In many organizations, the focus currently lies on the challenge to chart relevant customer behavior and its consequences as richly as possible, and to steer them in desired directions. This is the core of *Social Business Analytics*, the main theme of the third part of this book.

Part II

Your Big Data Potential

The Art of the Possible

1 “The Art of the Possible”

Intensive focus on business, organization and technology

In the nineteenth century, the German statesman Otto von Bismarck referred to politics as “the art of the possible.” This also applies to Big Data: operating cautiously, while simultaneously pulling out all the stops with the aim of maximizing results, clarifying decision-making, and stimulating new insight. The seven conclusions and recommendations presented at the end of this part dovetail perfectly with this aim. Here is a brief summary:

Big Data is the new, intensive, organization-wide focus on business, organization and technology. Accordingly, you must ensure that the organization has sufficient technological and analytical expertise, as well as the appropriate digital and organizational competences. After all, your aim is also to excel digitally and operationally. This is possible because making the best of Big Data in a structural way is becoming increasingly affordable.
With regard to Business Intelligence, *Data Discovery* is the next phase. It helps you combine lucid “magic moments” in your business operations with a significantly better performance through interactive visualization, exploration, planning and execution.

To start with, you can fire your imagination in inspiration sessions, followed by one or more concretization workshops in which you determine, in conjunction with an organization-wide team, where lucrative Big Data initiatives can be developed to suit your situation.

We analyze this constellation of potential in this part by distilling it into ten questions. We shall deal with them shortly, one by one. But first of all, we need to understand the importance of keeping an open mind.

Not science fiction but science facts

Charlie Beck, police commissioner in Los Angeles, is clear about the goal: there was no more money available, and no more police officers, so creativity simply had to increase. As the result of a political choice, all the stops had to be pulled out and results had to improve. For that reason the police force began to use a Predictive Policing algorithm, a Big Data solution that is now being applied in more than ten major American cities. In Los Angeles, crime decreased by 13 percent, and in Santa Cruz, which is also in California, by 26 percent.

The analytical software was developed by two mathematicians, an anthropologist and a criminologist, and is founded upon a model that predicts the aftershocks of earthquakes. For example, a criminal often returns to the scene of a previously committed crime. Such *aftercrimes* follow the same pattern as the aftershocks of an earthquake. On the basis of location, time, and type of crime, the software is capable of defining "prediction boxes" of 500 square meters.

This resembles science fiction, but is actually science fact. Historical facts and real-time events contain crosslinks and correlations of which we were unaware until Big Data came along. The retail trade works in a similar way. The case of the Target supermarket chain is iconic in that respect, continually growing its sales by perfecting the art of targeting. For example, data scientist Andrew Pole was able to predict whether or not a consumer was pregnant and the likely date of birth, on the basis of product purchases. This kind of predictive skill lies behind the Big Data potential that everyone is talking about these days. Take the Fraunhofer IAIS, for instance, the German institute for Intelligent Analysis and Information Systems. After extensive investigation on the basis of desk research, expert workshops and a survey, this Institute presented the Big Data innovation potential for German enterprises in March 2013 (see Question 2).

Big Data is your new colleague

Capital One Labs, a part of the Capital One bank and credit card company, is one of the many organizations that recruit Big Data talent. On Kaggle, a platform where many data professionals convene, we observed a Capital One job advertisement for this kind of data scientist. Anyone who applies for a job with Capital One Labs knows that he/she may be entering a Silicon-Valley-like culture, and will be given the following extraordinary assignment:

> *"to push the envelope to explore The Art of the Possible"*

In order to tap into the potential of Big Data, people have to understand the art of exploring the possible and of making the apparently impossible possible. ING bank articulates similar thoughts in its quest for new Big Data innovators: you have to stimulate out-of-the-box thinking, while showing that you "always behave in that way yourself," as the ING advertisement on the Monsterboard job vacancy site says (see Question 4). Big Data is your new colleague and the mission is to import and implement radical ideas: science fiction based on science facts. For instance, Agentschap Telecom is intensively engaged in the development of new methods, including techniques for tracing pirate radio stations. Johannes Brouwer, the head of the IT department there, states:

"The only restriction in the domain of Big Data is one's own imaginative power."

Spectacular results and promising experiments

Vigorous growth in turnover and spectacularly decreasing crime statistics – that is what the vast majority want to see. Of course, there are always restrictions – we shall mention many of them in the course of this part – but Big Data is genuinely "The Art of the Possible," and all new technologies make it possible to tackle matters in a radically different way.

Wal-Mart applies an intensive Big Data strategy by means of which it follows hundreds of millions of keywords and purchasing and searching behavior obtained via Google, Bing, Facebook, Groupon, Yelp, etc. These are imported via APIs after which the *bidding engine, analytics engine* and *pricing engine* are let loose upon them (see also walmartlabs.com/platform). Wal-Mart Labs, the "social data R&D" department, is seriously engaged in experimenting with, for example, semantic analyses. In this way, the so-called *Social Genome* is charted, consisting of detailed profiles of customers, topics, products, locations and events. The first results are now visible, such as the Wal-Mart app called Shoppycat, which presents gift ideas based on algorithms that interpret updates from social media. Presently, there are many such exciting new ideas and business cases.

Wait no longer: tap your Big Data potential

After the first phase of Big Data projects and experiments, the players are now beginning to seek more intensive contact with one another. To help organizations develop their data intelligence, dataset suppliers and analysis partners are active at opposite ends of the sector. Of course, the major IT partners are also involved. You yourself must determine with whom you wish to establish business connections, as well as the extent to which you wish to retain control of your Big Data activities in the long run.

MyBuys supplies both data and analytical capacity. The personalization engine of this young company is based on more than 200 million customer profiles and 100 terabytes of data in order to deliver real-time recommendations. MyBuys currently has more than 400 client organizations who are seeking to improve their sales. To financial service providers such as Capital One, which performs more than 80,000 Big Data experiments a year, there are many more possibilities based on aggregated transaction information, as long as they lie within the legal framework. The message to entrepreneurs is clear: seek a reliable partner and simply begin, according to the Fraunhofer IAIS, because:

"ranging from sensor data to Business Intelligence, from media analysis to visual information system, you are capable of doing more with data."

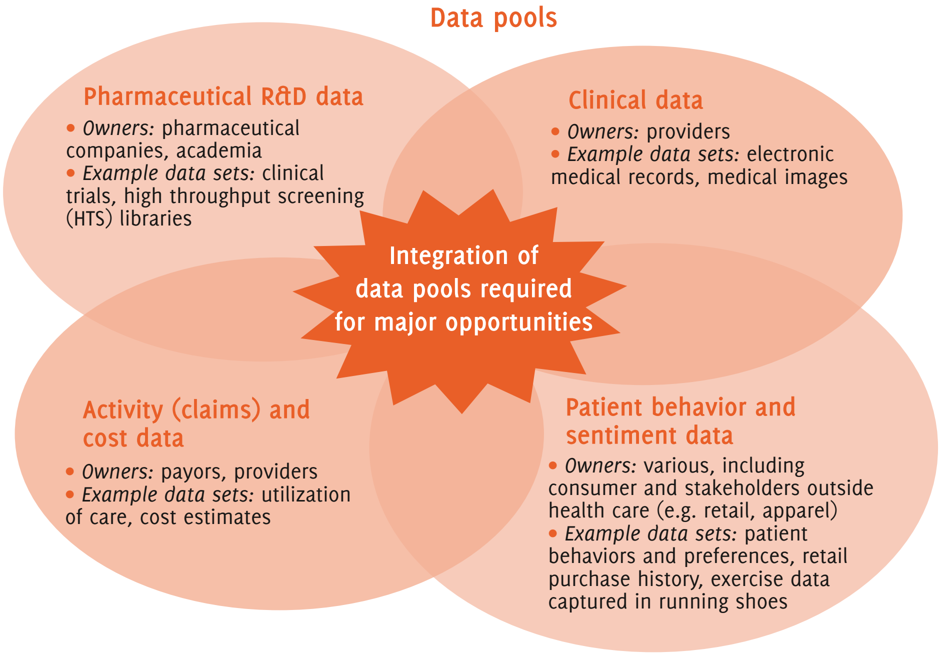
The core objective of all Big Data initiatives is to look beyond your own confines and to seek interesting combinations of internal and external, structured and unstructured data. This may be simply combining the current location of someone's telephone and transactions executed with his or her credit card. Too much distance between these may indicate fraud, certainly in combination with heightened transaction frequency. In this way, the data of telcos and banks could form a new kind of service. We do not explore all possible trans-sectorial connections in this part. Still, in concluding this introduction we need to emphasize that Big Data's potential can only be fulfilled when all intra- and inter-sectorial inefficiencies have vanished.

We shall take the healthcare sector as an example, but the energy or transport sector would be just as relevant. Healthcare provides many examples, such as Philips with its "hospital to home" strategy. Philips develops new products and services that give doctors, pharmacies, nursing personnel and even patients the opportunity to organize healthcare in a different way, with data technology and data visualization techniques.

Walgreens, the largest pharmacy in America, provides another example. It uses Big Data in linking point-of-sale data with social media, with data from customers' wearable computers, with data from partners such as the Nationwide Health Information Network, and with clinical data, all with the ultimate aim of improving patient care. Big Data potential radiates from the *data pools* that emerge from these kinds of new collaborations, as is shown by this overview of American healthcare presented by the McKinsey Global Institute:

The Big Data "Art of the Possible," based on mixing and matching, has demonstrably vast potential, but also requires much expertise, cooperation and coordination with regard to organization and levels of data and technology.

Four distinct big data pools exist in the US health care domain today with little overlap in ownership and low integration.

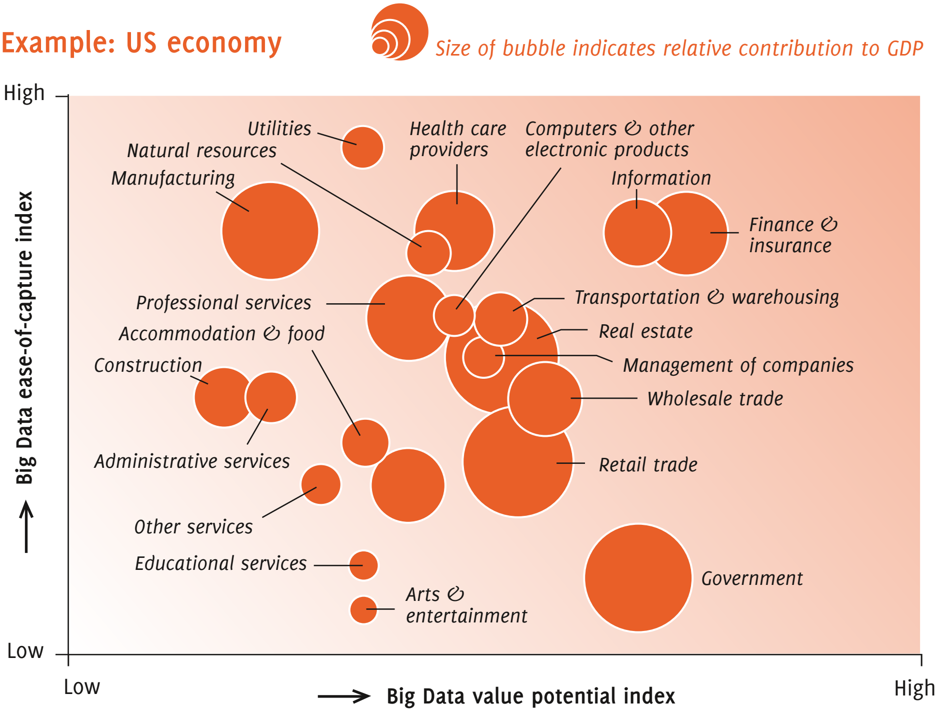

Data pools
Pharmaceutical R&D data
• Owners: pharmaceutical companies, academia
• Example data sets: clinical trials, high throughput screening (HTS) libraries
Clinical data
• Owners: providers
• Example data sets: electronic medical records, medical images
Integration of data pools required for major opportunities
Activity (claims) and cost data
• Owners: payors, providers
• Example data sets: utilization of care, cost estimates
Patient behavior and sentiment data
• Owners: various, including consumer and stakeholders outside health care (e.g. retail, apparel)
• Example data sets: patient behaviors and preferences, retail purchase history, exercise data captured in running shoes

Source: McKinsey Global Institute (2011)

2 Your Potential: the tension between *Could Be* and *Is*

After the iconic Big Data bubble chart (*Value Potential* versus *Ease of Capturing*) by means of which the McKinsey Global Institute predicted, in 2012, hundreds of billions of dollars in yields for various sectors of the American economy, there followed a whole procession of case descriptions that heralded a new data-intensive era. Potential applications were served up, technology was available, and the first smartly applied Big Data initiatives were already proving their productivity. Organizations were convinced: the status of Big Data as *The next frontier for innovation, competition and productivity* – the report with which the McKinsey Global Institute kicked off the new trend in 2011 – seemed indisputable. Perhaps there were no best practices to follow at that time, but "emerging next practices" had already left their mark, as Michael Chui of McKinsey remarked at the MIT Sloan CIO Symposium precisely a year later.

The ease of capturing Big Data's value, and the magnitude of its potential, vary across sectors.



Source: McKinsey Center for Business Technology (2012)

We are now a year further along, and the next step is, of course, to consider how Big Data's potential is currently being realized and how far we have advanced in this direction. At the end of 2012 and the beginning of 2013, a number of interesting studies were performed on ways in which organizations were engaged in developing their Big Data potential. On the basis of these studies, the German Fraunhofer IAIS (the Institute for Intelligent Analysis and Information Systems), TCS (Tata Consultancy Services) and SAS, in conjunction with the CMO (Chief Marketing Officer) Council, responded to the question about the potential of Big Data, approaching the question from several complementary angles.

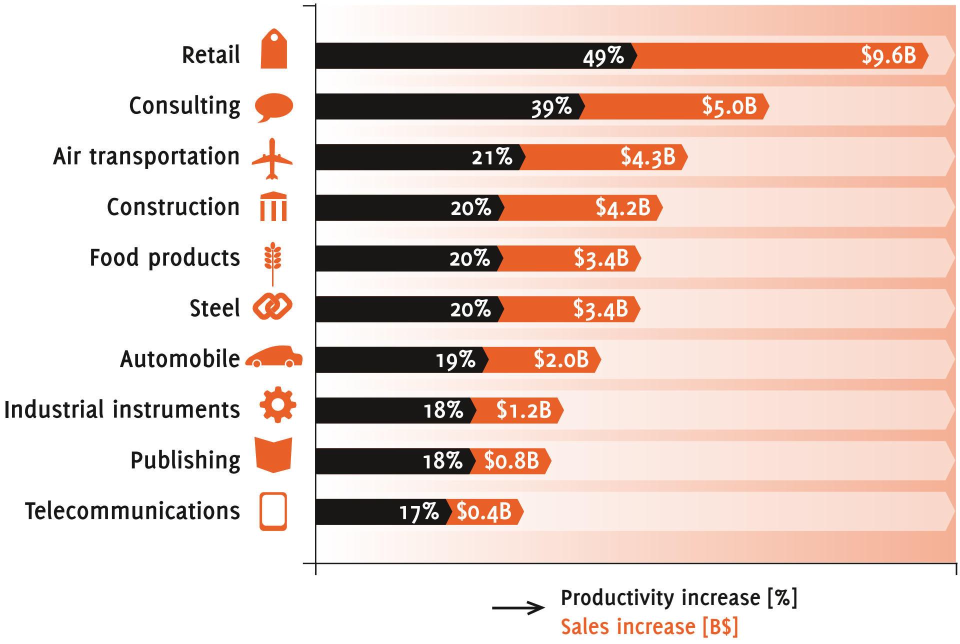
They give new insights into the way in which the ten simple questions (which are presented once more at the end of this section) can be answered. They thus chart the tension between *Could Be* and *Is* – thus, your potential. On every occasion, in order to make well-considered decisions, you must measure your own position and ambitions against your potential and the concrete lessons you've learned.

You should not abandon past data-related achievements. On the contrary, you can examine them more closely, broaden them, expand them and especially integrate them. But you should only begin this integration with a good knowledge of details – of developments in technological and organizational fields, drivers, needs, restrictions, requirements and impact. Act according to your best insight with a lucid priority plan, at your own speed and with clear aims and performance indicators. You may decide to disregard this kind of general advice in favor of addressing these details in a concrete way and, above all, by being honest about the situation throughout the organization and involving all stakeholders.

Accordingly, this part is no simple roadmap but it is a concrete and solid conclusion built around ten straightforward core questions. We envisage Big Data integration as the basis of new Business Intelligence; we see technology that enables a data-intensive approach to the issues touched upon in the questions; we examine your plan to become data-driven; we analyze your degree of Digital Advantage, and we conclude with a forecast of Big Data potential in 2020. Now is the time to develop your Big Data potential, and that is why you will find a checklist of twenty items at the end of this part.

Our knowledge and experience with what we call Big Data – immense amounts of data, and/or very varied data, and/or very rapidly changing data – is expanding day by day. Originally the notion of Big Data was linked to *Volume*, *Variety* and *Velocity*, but another three Vs have now joined up: *Veracity*, *Variability* and, of course, *Value*. The last three form the universal yardstick against which we must measure all data-intensive activities because, quoting Tom Davenport the analytics guru, we wish to use them to move from *descriptive* to *predictive* and ultimately *prescriptive analytics*. If the predictive analyses turn out to be right, we will win time and effectiveness by immediately prescribing and adhering to them.

Big Data is thus a wake-up call to use all conceivable data that we have at our disposal in order to become genuinely "data-driven." The following figure shows the *Value* component of effective data usage, presented for ten different business sectors.



Source: University of Texas (2011)

Isn't it marvelous? But how do you do it? The website http://UnlockingBigData.com offers a simple Big Data self-test, and this should supply all the necessary answers. Just answer the questions in the categories of *Data, People, Technology, Process* and *Intent*, and you will have a snapshot of how "mature" your business is, with epithets such as *Novice, Beginner, Competent, Proficient* and *Expert*. Good to hear, but you can probably already state with a reasonable degree of accuracy just where you are located on the Big Data scale.

In question 8, about what is looming on the horizon, we present the results of a much-renowned study by MIT in conjunction with Capgemini. The conclusion is that a digital strategy does pay off. Companies that invest more in digital and organizational effort have a higher turnover, more margin, and better market value. Big Data is the latest new development in the context of your digital strategy.

Removing ROI (Return on Investment) from your Big Data activities, closing the gap between potential and reality, begins with posing the right questions. Many questions have already been dealt with in other parts of this book. You can consult them again to gain some inspiration. Thispart lists the following ten most important questions in order to help formulate a concrete plan. We offer a number of key points, by means of which you can draw your own roadmap.

In technological terms, questions 6 and 7 pinpoint the difference with traditional RDBMS environments, but a purposeful data-focus on the combination of business, organization and technology is the core as well as the aim of the exercise.

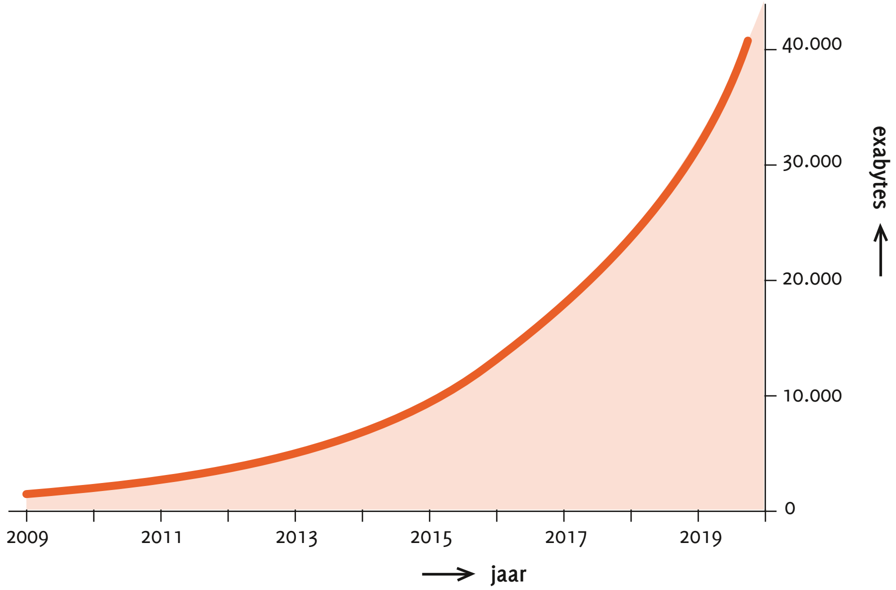
Question 1 Why Big Data intelligence?
Question 2 What new insights can I expect?
Question 3 How will these insights help me?
Question 4 What skills do I need?
Question 5 How do Big Data pioneers organize data management and IT processes?
Question 6 How can I merge my structured and unstructured data?
Question 7 Which new technologies should I be watching?
Question 8 What is looming on the horizon?
Question 9 What does this mean in organizational terms?
Question 10 How does this affect everyday life?
(please see our conclusion and checklist)

Question 1: Why Big Data intelligence?

Answer: Because large amounts of data become available at little cost. These data contain valuable insights and their processing is attainable in technological and analytical terms.

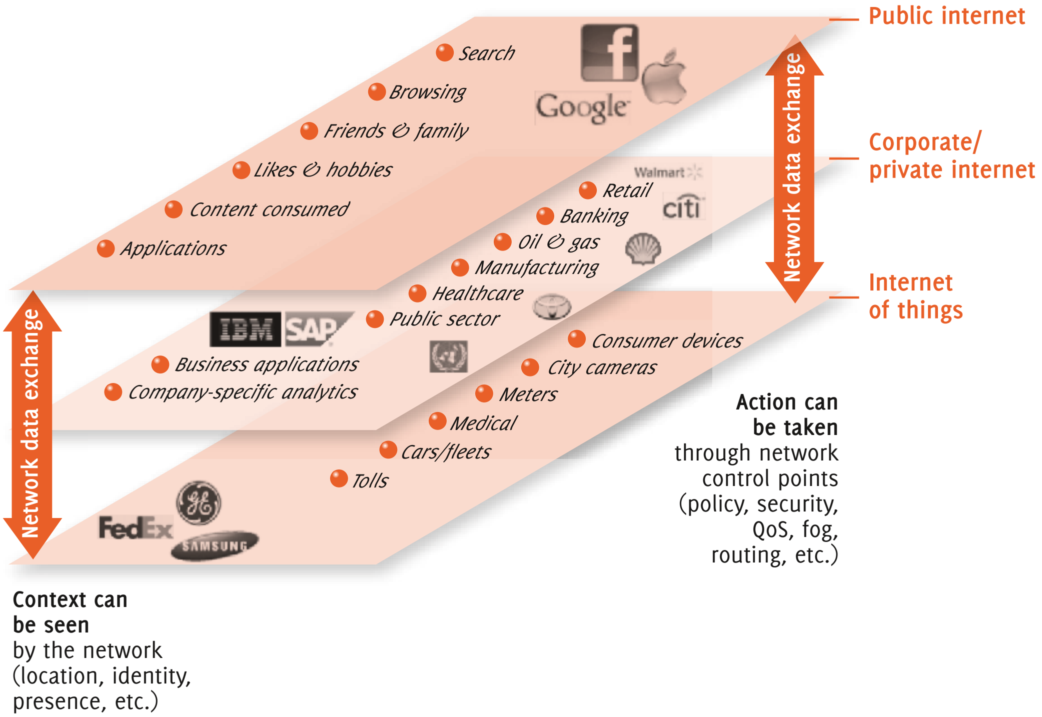
In 1965, Gordon Moore made a prediction in the *Electronics Magazine* that the number of transistors in a chip would double every two years. Until the present day, that prediction has been true every year, and it is now referred to as “Moore’s Law.” The two-year constant has now been adjusted to eighteen months. In the past thirty years, this exponential growth has become visible not only in the mathematical power of processors but also in working memory, storage space, bandwidth, the number of electronic sensors, and the quantity of data.

In its study entitled *The Digital Universe in 2020*, IDC calculated that the second decade of this century would witness a growth in the quantity of data to 40,000 exabytes, in other words: 40 billion gigabytes. This is equivalent to 5,200 gigabytes for each person on Earth. In 2005, the counter indicated only 130 exabytes.



In this decade, the amount of data worldwide will grow to 40,000 exabytes
Source: IDC & EMC (2012)

According to IDC, a quarter of the current Big Data mountain contains information that is useful for analysis. In 2020 this will grow to one third. Therefore we will be dealing with exabytes of valuable data in the coming years. In its working paper entitled *Defining a Taxonomy of Enterprise Data Growth*, the Center for Large Scale Data Systems concluded that most data that are currently being created within organizations are unstructured data. The data originate in a combination of three internet types: the Public Internet, the Internet of Organizations, and the Internet of Things.



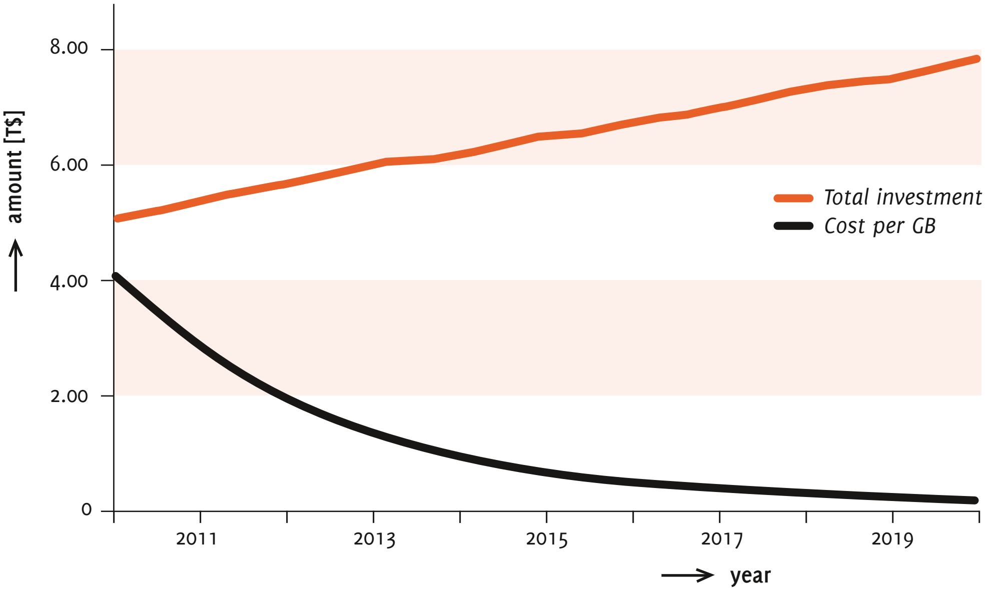
Source: Cisco IBSG (2012)

The current decade will largely be concerned with the question of how organizations will tap into this immense Big Data potential and what exactly that will mean. MIT professor Erik Brynjolfsson compares it to the invention of the microscope, which also enabled great surprises and new insights in many fields and disciplines:

> *"The microscope made it possible to see and measure things in a way that was unprecedented. Now we are dealing with the modern equivalent of the microscope."*

Thanks to advanced hardware and software, we are now capable of zooming in and out at great speed, with the aim of discovering structures and links that will give us more insight and enable us to make better decisions and find more effective solutions. Marketing, healthcare, energy supplies, transport, every type of service provision – in short, all forms of applied science take on a completely new aspect, and this will have an unprecedented impact.

This will also bring opportunities for the IT sector itself, which will also undergo transformation. Investments in IT will rise thanks to the need for Big Data insights, while the costs per gigabyte will decrease from four dollars to a few cents, primarily due to the explosive increase in data.

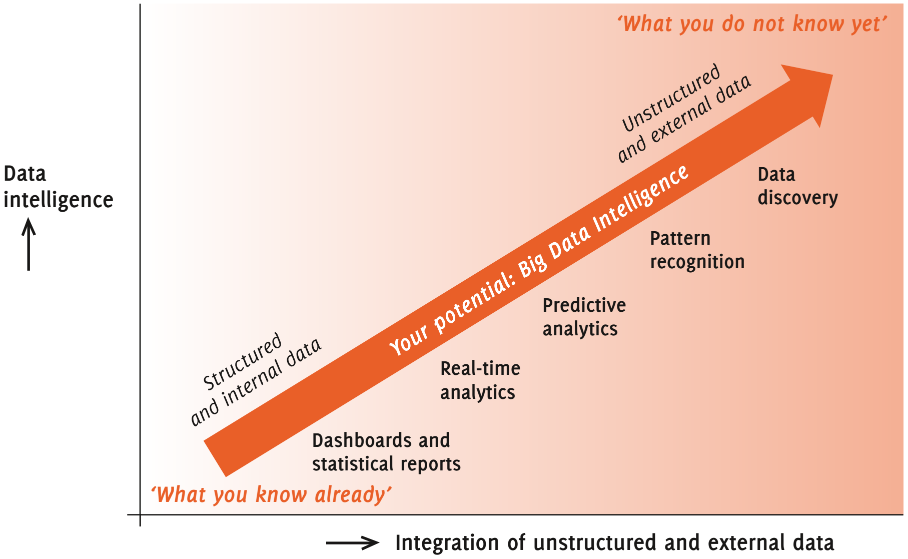


In 2020 one gigabyte will cost almost nothing and investments will rise substantially
Source: IDC & EMC (2012)

This picture is confirmed by the results of a study entitled *Big Data: Big Potential, Big Priority* that Cisco published at the end of March 2013. It presented the results of a survey on the status of Big Data in eighteen countries. More than half of the companies expected to make higher investments in the coming years thanks to the Big Data priorities that are now being specified.

Potential lies in unstructured data and business transformation

Capitalizing on Big Data potential goes hand-in-hand with investment aimed at getting hold of unstructured and external data: begin on a small scale and subsequently build up the capacity to harvest the unstructured and external data. That is the challenge, that is the Big Data potential that is consistently mentioned: accessing unstructured and external data to develop new insights. That is, concisely summarized, what the term "next-gen Business Intelligence" refers to. It is ultimately a question of "data intelligence." It represents a whole new generation of approaches, tools, insights, and different ways of working (faster, better and much more efficient).


'What you do not know yet'
Unstructured and external data
Data intelligence
Data discovery
Pattern recognition
Your potential: Big Data Intelligence
Predictive analytics
Real-time analytics
Structured and internal data
Dashboards and statistical reports
'What you know already'
Integration of unstructured and external data

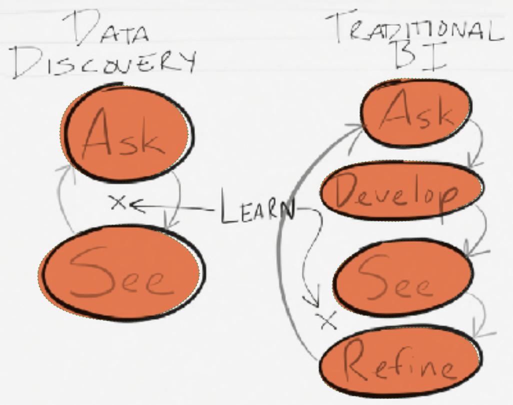
Business Intelligence on the basis of Big Data integration leads to new insights, and faster and better decision-making

In the meantime, the flood of success stories continues to grow. Shell detects oil fields through Big Data, the Netherlands Forensic Institute investigates data to trace criminal behavior, the Telecom Agency monitors social media to protect the airwaves, and KLM uses Big Data to gain better insight into the search behavior of customers. But even here in the Netherlands, capitalization on and acceptance of Big Data are still in their infancy.

Nevertheless, data integration and Big Data are already familiar faces in some organizations and among individual experts, think tanks, and particularly in more scientific and data-intensive organizations such as the Netherlands Forensic Institute, academic hospitals, telecom companies and the Telecom Agency, banks, insurers, credit card companies and energy suppliers. In such settings, Big Data technology complements traditional research methods such as data-warehousing, data-mining and Business Intelligence. In several cases, in-house experiments are carried out with new technologies, and the impact of the new technology on the existing operational landscape is closely scrutinized.

Nowadays there is a much greater volume of data available from customers and competitors than was the case roughly five years ago. And much greater data volume is certainly coming. Fortunately, new technology, such as Hadoop and NoSQL ("Not Only SQL"), is already capable of coping with growth. This development brings many new opportunities, such as the upsurge of the so-called *Data Discovery*, which enables an interactive, visual, rapid-cycle question-and-answer analysis on the basis of modern data technology, enabling more rapid insight.

Capital One, the fifth largest credit card company in America, indicates that 80,000 data trials are performed annually. In this way, important new possibilities are discovered and the ROI of every marketing dollar is maximized. (See an up-to-date overview of Data Discovery tools on the website of Applied Data Labs.)



Data Discovery is the next-generation Business Intelligence

In terms of technology, the potential of Big Data lies in its capacity to interpret and apply external and unstructured data. In terms of organization, we are dealing with a major transformation. The Big Data potential of every organization is developed along these two axes – the capacity to process unstructured data and the capacity to transform your organization.

In its report, McKinsey explicitly accentuated the necessity of transformation. Without this transformation, there can be no mention of efficiency. This is much more than simple rhetoric. The expected leverage effect of Big Data is described by Viktor Mayer-Schönberger, professor at the Oxford Internet Institute, and Kenneth Cukier, Data Editor of *The Economist*, in their latest book:

> *”[Big Data] refers to the things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value in ways that change markets, organizations, the relationships between citizens and governments, and more.”*

Cukier and Mayer-Schönberger's essay "The Rise of Big Data" was featured in the April/May 2013 edition of *Foreign Affairs*, the leading American magazine on international relations. This demonstrates just how topical and critical Big Data and its potential applications are, globally. This engagement is expected to continue into the foreseeable future.

POLE POSITION: A CONVERSATION WITH RADEK SIKORSKI
FOREIGN AFFAIRS
MAY/JUNE 2013
The Rise of Big Data
Kenneth Cukier & Viktor Mayer-Schoenberger
Austerity Bites
Mark Blyth
The Irony of American Strategy
Richard Haass
Africa's Boom
Shantayanan Devarajan & Wolfgang Fengler

Question 2: What new insights can I expect?

Answer: Two studies, one by Fraunhofer in Germany and the other by TCS, indicate where the greatest global potential of Big Data lies at present.

Big Data – Advantage through knowledge: Innovation potential analysis 2013

In March 2013, the German Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) and the Ministry of Economic Affairs and Technology revealed how organizations in Germany, the fourth largest economy in the world, currently regard the innovation potential of Big Data. IAIS is a part of the Fraunhofer organization for applied research and, with more than 20,000 employees in Europe, it is the largest institute of its kind. The Institute categorizes the present use of Big Data as follows:

1. Monitoring markets to identify sales opportunities
2. Creating personalized offers
3. Recognizing repeat customers at an early stage (*churn*)
4. Recruiting staff
5. Forecasting sales, for planning and control
6. Predicting wear and maintenance
7. Directing management, decision-making & control
8. Detecting fraud
9. Estimating financial risks

10. Recognizing cyber attacks
11. Improving products
12. Developing innovative products

The *Innovation Potential Analysis 2013* formulated by Fraunhofer IAIS comprises three sections: an inventory based on theoretical research, five sectorial expert workshops, and an online survey. This last element produced the following seven main results:

1. In general, trade & industry are aware of the possibilities that Big Data analysis offers.
2. The greatest potential lies in building up a strategic competitive advantage (69 percent), followed by a higher turnover (61 percent) and cost-savings (55 percent). Higher productivity and data-driven planning and decision-making are also mentioned as objectives.
3. Not one but several organizational departments want to benefit from the advantages provided by Big Data, and a large group mentioned management in particular.
4. Marketing and operational aims score highly, in contrast to aims relating to IT.
5. At this moment, organizations are not optimally configured for Big Data analysis. The budget is too small, responsibilities and governance are not clearly specified, and competency needed to handle Big Data efficiently is inadequately developed.
6. Various issues will be dealt with in the coming five years, and no one anticipates a decline in the Big Data budget.
7. In order to be able to master the required competences, most respondents would like to learn about the best practices and are willing to undergo training.

The five sectorial expert workshops

Expert workshops were held for the following sectors: financial service provision, telecom and media, insurers, trade and retail, and market research. The aim was to generate a roadmap for the future. To start with, it is evident that Big Data is not exclusively an IT theme, and it does serve strategic aims. Moreover, it is also clear that interorganizational and intersectorial cross-fertilization, as well as a combination of findings and points of concern, may constitute the next major Big Data development, ultimately resulting in new and more complex business models.

Financial service provision

In the short term, financial service provision is primarily interested in efficiency: costs set against results. In this context, too, companies are fully aware of the fact that new providers, from other sectors for example, could develop new products on the basis of

Big Data – products that are more finely tuned to the individual customer. Everyone is aware of the innovation potential that lies dormant in Big Data, but it is unclear who is going to be the first to actually make use of this.

Telecommunications and media

The telecommunications and media sectors have at their disposal a huge volume of data that enables them to provide better service to consumers. However, the challenge is to capitalize fully on this kind of information. Also, this framework demands more efficient internal processes. In the longer term, the automation of data analysis will lead to new and more complex business models.

Insurers

Experts in the insurance sector recognize the advent of current and future Big Data challenges. At present, handling personal data is governed by law on the one hand, and through consent from the individual on the other. In the future, other data sources may arise (as a result of monitoring and predictability, for instance), that could lead to the development of behavior-related insurance products. In this context we can think of data from telematics systems, social media, and electronic health records. Whether or not consumers support such innovative uses of data, and whether or not rules and regulations will be adjusted to respond, remains an open question.

Trade and retail

In trade and retail, Big Data analysis is clearly regarded as an opportunity to modernize. The infrastructure for data-intensive services will increasingly be offered as a resource. As a consequence, smaller organizations will also be able to benefit from Big Data with respect to segmenting and targeting. In the longer term, the data will become a product.

Market research

Market researchers, whether as a separate sector or as a subsector in larger organizations such as the credit-card industry or other financial or telecom-oriented service providers, are ideally suited to capitalize on the opportunities offered by Big Data. Algorithms, statistics and advanced visualization can be applied in order to present up-to-date and detailed market analyses and opportunities at any given moment. The development of this kind of data science is already rapidly rising, although data scientists are still very scarce. Fraunhofer IAIS is currently actively engaged in this project and offers a data science curriculum for data analysts, business analysts, marketers and financial personnel.

The emerging big returns on Big Data 2013

In December 2012 and January 2013, TCS held a worldwide survey among 1,217 major organizations and published the findings in a report entitled *The Emerging Big Returns on Big Data*. The current overall picture of the field of Big Data can be summarized in the following five points. They represent the first stage of Big Data analysis in 634 organizations worldwide, over 12 economic sectors of which 83 percent have a turnover of more than 1 billion dollars.

1. More than half of the major organizations surveyed indicated that they developed Big Data initiatives in 2012.
2. Of these, 80 percent had improved business decisions as a result.
3. The majority of the data used remains structured, and comes from internal sources.
4. The greatest challenge within major organizations is ensuring that existing silos share their data.
5. In 2012, the median investment made by large organizations (not the average value, due to several substantial exceptions) was 10 million dollars.

We recognize that new Big Data-based insights can emerge in many areas of an organization: sales, logistics, marketing, HR and customer service. Expectations are highest for what new insights Big Data can bring to key concerns:

1. Identification of consumers who can generate most value.
2. Ability to monitor the quality of one's own products in detail.
3. Discovery of new needs among consumers, needs that current products cannot meet.

Question 3: How will these insights help me?

Answer: There are already concrete ROI expectations for eight corporate functions, and the focus and value of each function can be specified.

The TCS study predicted what the new insights will deliver. The results are set out below. We can easily recognize the twelve Big Data areas of the Fraunhofer IAIS, which we listed earlier in this part.

Corporate function	Highest value	Second-highest value	Estimated ROI over 2012
Logistics	Monitoring product shipment	Identifying peaks in costs	78%
Finance	Risk measurement	Forecasting	69%
Customer Service	Identifying consumer loyalty	Website behavior	56%
Sales	Identifying the most important consumers	Spotting cross-selling opportunities	54%
R&D	Monitoring product quality	Spotting new consumer needs	48%
HR	Estimating employee retention	Effectiveness of recruitment	48%
Operations	Product tracking	Planning deliveries	42%
Marketing	Measuring campaign effectiveness	Measuring media effectiveness	41%

Where can we expect a positive ROI with regard to Big Data?

Please note that an ROI approach is not always the best start to a Big Data project, especially if you wish to promote Data Discovery.

Question 4: What skills do I need?

Answer: BI professionals, marketers, sales people and IT management must collaborate with data scientists to fully master Big Data. Therefore recruitment and training are both essential.

In his 2012 book *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*, Bill Franks explains how organizations should set up and implement Big Data initiatives incrementally. Franks is the Chief Analytics Officer at Teradata, and also director of the Business Analytic Innovation Center of Teradata and SAS. Harvesting Big Data, as described in the figure on page 15, moves from internal to external and unstructured data in small increments. According to Franks, and many other Big Data experts agree with him, the quantity and the quality of insights gained will thus increase exponentially. He rightly emphasizes that an overall strategy is extremely practical and necessary to access the latent potential of Big Data. It is not only a technological issue – how do you blend structured and unstructured data? – but certainly an organizational issue as well: how do you transform the organization? Franks comments:

> *"The biggest value in Big Data can be driven by combining Big Data with other corporate data. By putting what is found in Big Data in a larger context, the quantity and quality of insights will increase exponentially. This is why Big Data needs to be folded into an overall data strategy as opposed to having a stand-alone Big Data strategy."*

Franks emphasizes the paradox that, although organizations in traditionally data-analytical economic sectors can choose from a greater pool of suitable new employees, extra effort is nevertheless needed to keep up with the competition, due to that sectorial lead. Organizations in sectors without a data-intensive culture can gain a lead over their rivals though new analytical initiatives, but they may also have to reinvent the wheel themselves in some situations. In addition, several studies over the past few years have shown that the demand for Big Data analysts and data scientists easily exceeds the supply. As an illustration, here is a profile of a data scientist, so you can see just what kind of creature we're talking about.

Vacancy for Data Scientist on Monsterboard, April 2013

ING bank is seeking an innovative, creative and enthusiastic data scientist. Someone who has no difficulty with tackling and solving complex marketing issues with the aid of substantial amounts of data (Big Data) and modern IT resources. Someone who can stimulate and inspire colleagues to think "out of the box" and is himself/herself a classic example of this approach. […] This data-driven marketing makes use of a voluminous data and application infrastructure that is currently in the process of being completely renovated. In this framework, the focus lies on the establishment of a real-time, predictive infrastructure and a different way of (co-)operation. Aspects that are now at the forefront: Big Data, Netezza, Hadoop, Agile.

Function requirements:

The chosen candidate will be an academic with an informatics/mathematics/econometrics background, preferably with the accent on machine learning or artificial intelligence. The candidate may be currently busy with a PhD in one of these disciplines. Candidates should be familiar with (and, if possible, have experience with) technology and methodology in the field of structured and unstructured data (Netezza, Hadoop, MapReduce, Hive, R, SAS). They possess demonstrable knowledge and experience with statistical and mathematical methods of data analysis (such as neural networks, random forest, text-mining for example) and their application in business processes. In addition, candidates have roughly the following profile:

- *Are extremely inquisitive and continually searching for explanations*
- *Can cope comfortably with setbacks, are true stayers.*
- *Present sharp analyses – concise, to the point, convincing.*
- *Maintain a clear overview in complex, obscure situations.*
- *Are good in communicating, presenting and selling ideas.*
- *Are enthusiastic and genuine team players.*
- *Are quick to absorb new techniques and software.*

It is not our intention to place all the work on the shoulders of Big Data scientists. BI professionals, marketers, sales people and IT management must work with data scientists on new competences in order to fully master the complexities of Big Data.

The themes of sharing information and new technology are prevalent. In this setting, features such as expertise, including the ability to apply insight and prioritize, and willingness to contribute are major determinants. Attention must also be paid to developing trust between data scientists and business managers. Good data visualization is also important. Marketing and data experts want a clear business focus on the investments in Big Data, as well as on the various sorts of data that are needed for diverse issues. This is directly related to the previously mentioned availability of data from the silos in the business units.

Question 5: How do Big Data pioneers organize data management and IT processes?

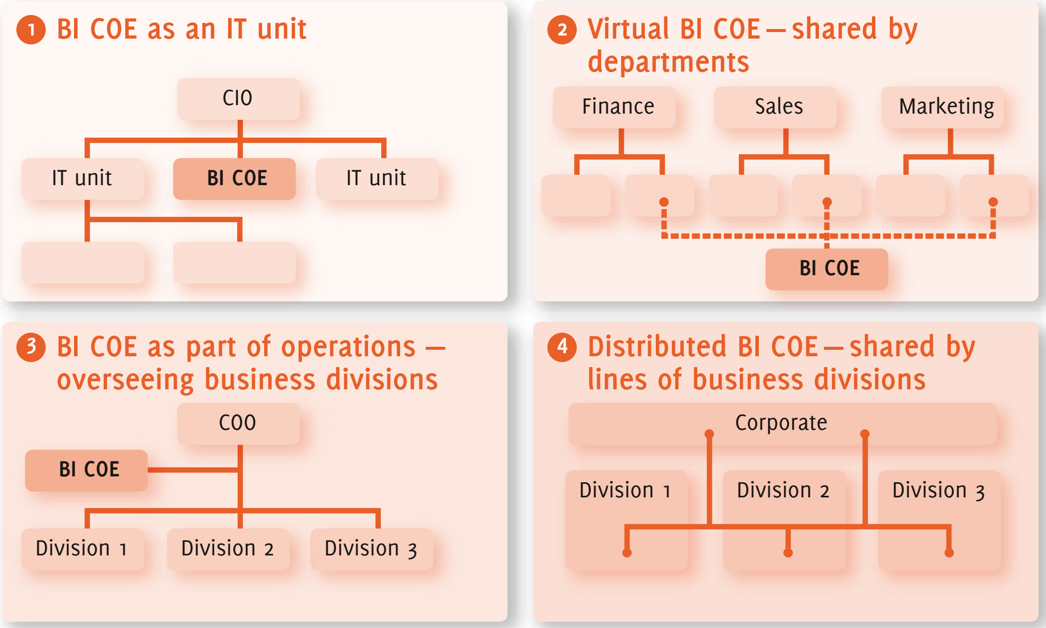
Answer: There are various ways to support the Big Data project; for example, by means of a so-called "next-generation" BI CoE (Center of Excellence). The most important way to provide support is by building successful collaboration between CMO, CIO and CDO (Chief Data Officer, if present), as is the case with Starbucks.

There is much commotion around the way in which organizations should shape their "real-time fact-based decision-making." Getting good results requires more than merely implementing technology such as Hadoop, Oracle Exadata, SAP HANA, SAS/Teradata or IBM BigInsights. To start with, there must be change at the top. Anyone aspiring to capitalize on the potential of Big Data will have to pull some strings at the executive level. Depending on the source, the new executive function will be filled by a Chief Data Officer, Chief Decision Officer or the all-embracing Chief Digital Officer. What a Chief Data Officer and his/her Data Management Office actually do is reported in *The Role of the Chief Data Officer in Financial Services* by Capgemini. Topping the list of eighteen reasons to appoint a CDO, numbers 1, 2 and 3 mention the silo problem.

To generate good decision-making, business analysts must have various sorts of data at their disposal: data from the data warehouse, but also data from the Web, Big Data and information from production systems, as well as input from partners and suppliers. Analysts spend more than half of their time acquiring this data. Because we generally wish to examine the facts as closely as possible, less time remains for solid analysis. For this reason, it is advisable to deliberate seriously and to continually evaluate the design and delivery of effective next-generation BI. Tools, datasets, knowledge and skills, and the right mindset in the organization are essential components of such structures.

It is common to concentrate the new core expertise in a (next-generation) Business Intelligence Center of Excellence (BI CoE), also referred to as a Business Intelligence Competence Center (BICC), and to coordinate it from this source. There are four main variants that are applied with businessmen, IT people and data experts, either separately or in a certain combination for a fixed time:

1. an IT unit that reports to the CIO
2. a virtual CoE, present in different functional business units
3. an operating unit that reports to the COO
4. a distributed CoE, present in the divisions and the overarching corporate organization.


1 BI COE as an IT unit
CIO
IT unit
BI COE
IT unit
2 Virtual BI COE – shared by departments
Finance
Sales
Marketing
BI COE
3 BI COE as part of operations – overseeing business divisions
COO
BI COE
Division 1
Division 2
Division 3
4 Distributed BI COE – shared by lines of business divisions
Corporate
Division 1
Division 2
Division 3

Four ways to embed a next-gen BI CoE in the organization
Source: Kalakota (2012)

Please note: It is self-evident that the virtual CoE and the distributed variant will report to the CDO. Of course there must be good connections with the CIO office, which may be formally established without there necessarily being two captains on one ship or any inevitable clashes. The CIO office will be more engaged with the backend of the information systems, while the CDO and the CMO and their staff will explicitly direct their efforts to the acquisition of value from data for the consumer and the organization.

A remarkable work in this context is the study entitled *Big Data's Biggest Role: Aligning the CMO & CIO – Greater Partnership Drives Enterprise-Wide Customer Centricity* by SAS and the CMO Council, which was published in March 2013. Due to the increasing focus on Big Data analysis, the CMO and the CIO Office are now growing toward one another, whereas a wide gap was clearly evident not so very long ago.

Other handy tips for steering the proliferation of data and the necessary data integration in the right direction (in short: how we design the next-gen BI) are presented in the report *Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO*, by IDC and SAS. (In the figure below we have streamlined the phrasing and the functional elements in the report.)

	Maturity			
	Stage 1	Stage 2	Stage 3	Stage 4
Data Governance	Little or none (Skunk words)	Initial data warehouse model and architecture	Data definitions and models standardized	Clear master data management strategy
Technology & Tools	Simple historical BI reporting and dashboards	Data warehouse implemented, broad usage of BI tools, limited analytical data marts	In database mining, and limited usage of parallel processing and analytical appliance	Widespread adoption of appliance for multiple workloads Architecture and governance for emerging technologies
Staff Skills (IT)	Little or no expertise in analytics – basic knowledge of BI tools	Data warehouse team focused on performance, availability, and security	Advanced data modelers and stewards key part of the IT department	Business Analytics Competency Centre (BACC) that includes "data scientists"
Staff Skills (Business/IT)	Functional knowledge of BI tools	Few business analysts – limited usage of advanced analytics	Savvy analytical modelers and statisticians utilized	Complex problem solving integrated into Business Analytics Competency Centre (BACC)
Financial Impact	No substantial financial impact No ROI models in place	Certain revenue generating KPIs in place with ROI clearly understood	Significant revenue impact (measured and monitored on a regular basis)	Business strategy and competitive differentiation based on analytics

Loosely taken from IDC & SAS (2011)

The figure presents, moving from top to bottom, the coherence of adequate data governance, the right technology and tooling, and staffing in the IT and business/IT department. The consequence – the anticipated impact on the financial results of the organization as an accepted indicator of the overall business performance – is displayed at the bottom. The increase in maturity is shown in four stages from left to right, moving from minimum to optimum. It is up to you to use the information provided in this report to determine where your challenges and opportunities lie, and what is the most suitable route for you to take toward growth.

CIO and CDO a successful team at Starbucks

Organizations do differ. Big Data may provide an impulse for the CIO and the CMO to collaborate more intensively, but also a newly appointed CDO (Chief Data Officer) and the CIO can form an excellent team, as is the case at Starbucks.

Starbucks is worth 13.3 billion dollars, and is extremely successful with its combination of brand stores and digital ventures such as social media and mobile networks. A CDO was appointed in March 2012 and he now supervises 110 employees. He works in close cooperation with the CIO, who was appointed at the same time and whose office comprises 760 employees. The great change in the first year was the repositioning of technology and digital services by orienting them more toward the customers and the shops.

All Starbucks' digital projects have now been combined: web, mobile, social media, digital marketing, loyalty programs, e-commerce, wifi, the Starbucks network and new shopping technology. In bygone days, worldwide digital marketing, the Starbucks cards and mobile payments, and loyalty programs were three separate units. These have now been successfully integrated. The CIO and the CDO work in harmony and the entire digital Starbucks roadmap is thoroughly examined on a weekly basis. The traditional, separate Centers of Excellence have been replaced by intensive project teams, working in rapid iterations (*MIT Sloan Management Review*, 2013).

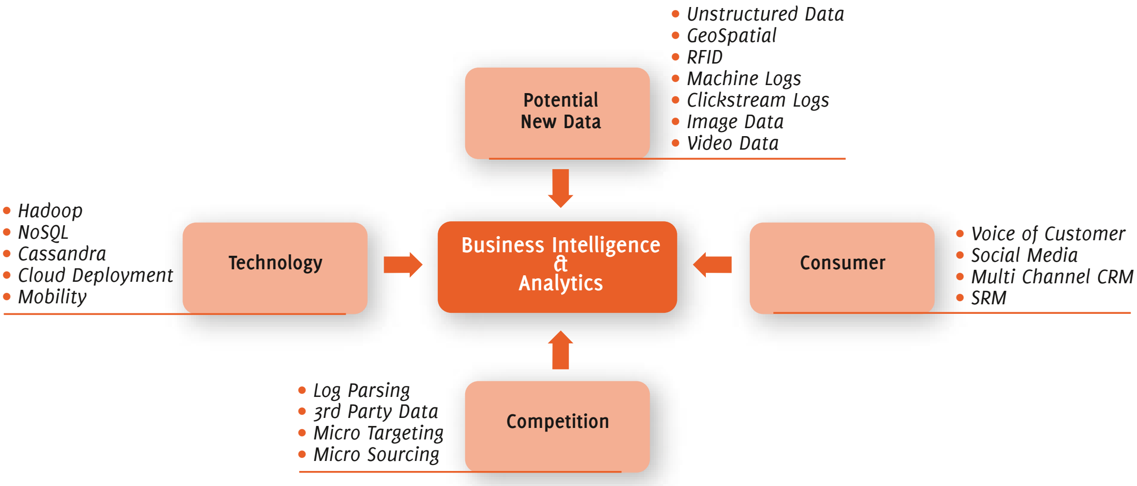
Question 6: How can I merge my structured and unstructured data?

Answer: Business Intelligence must become faster and more analytical. Ideally you would use a modern, integrated information architecture with specialist hardware, a high-speed network, and in-memory analytics for this purpose.

For ten years, stretching from 2002 to 2012, Sam Palmisano was CEO of IBM. In January 2010, he made his "Decade of Smart" speech. In the course of the second decade of the 21st century, Palmisano predicted that we will find an increasing number of answers in the rapidly growing digital dataflow that we have at our disposal.

Palmisano deliberately avoided talking about Big Data. All conceivable data taken together – large, small, structured or unstructured, relational or graph, metadata and masterdata, in ontologies and taxonomies – will jointly lead to new insights and better and quicker decisions.

Here too, the *devil lurks in the details*. In our humble assessment, everything that is currently referred to as "Business Intelligence" must dig a bit deeper or shift up a couple of gears. In other words, everything must become more analytical and faster. Krish Krishnan, who published *Building the Unstructured Data Warehouse* with data guru Bill Inmon at the beginning of 2011, and whose new book *Data Warehousing in* *the Age of Big Data* recently appeared, presents the new challenges in the field of BI schematically, as in the following figure.



Source: Krishnan (2012)

A more thorough analysis in the context of data integration will be presented shortly, but the impact will be evident. In his overview entitled "Big Data & Analytics," from which the above figure showing the data challenges to BI has been borrowed, Krishnan sums up the central developments, challenges, solutions and limitations that affect Business Intelligence in the framework of Big Data: the inclusion of all kinds of unstructured and external datasets in the analyses and insights for the organization. In its generality, Big Data is akin to the more traditional enterprise data domains in the organization, as is outlined in the following Oracle table:

Data Realm	Structure	Volume	Description	Examples
Master Data	Structured	Low	Enterprise-level data entities that are of strategic value to an organization. Typically non-volatile and non-transactional in nature.	Customer, product, supplier, and location/site
Transaction Data	Structured & Semi-structured	Medium-High	Business transactions that are captured during business operations and processes	Purchase records, inquiries, and payments
Reference Data	Structured & Semi-structured	Low-Medium	Internally managed or externally sourced facts to support an organization's ability to effectively process transactions, manage master data, and provide decision support capabilities.	Geo data & market data
Metadata	Structured	Low	Defined as "data about the data". Used as an abstraction layer for standardized descriptions and operations (e.g. integration, intelligence, services).	Data name, data dimensions/units, definition of a data entity, or a calculation formula of metrics
Analytical Data	Structured	Medium-High	Derivations of the business operation and transaction data used to satisfy reporting and analytical needs.	Data that reside in data warehouses, data marts, and other decision support applications
Documents and Content	Unstructured	Medium-High	Documents, digital images, geo-spatial data, and multi-media files.	Claim forms, medical images, maps, video files
Big Data	Structured, Semi-structured & Unstructured	High	Large datasets that are challenging to store, search, share, visualize, and analyze	User and machine-generated content through social media, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies, and genomics

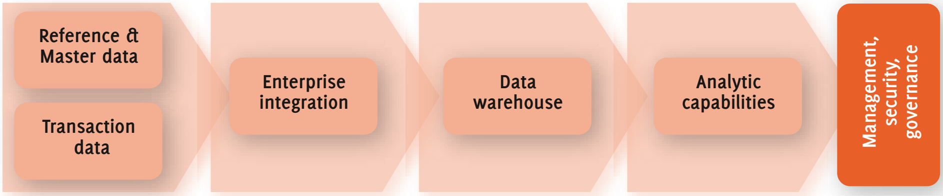
Source: Oracle (2012)

Altogether, this requires ongoing development of the information architecture, which we should envisage – according to the Oracle – in three steps:

Step 1

This is the traditional situation with only structured data:

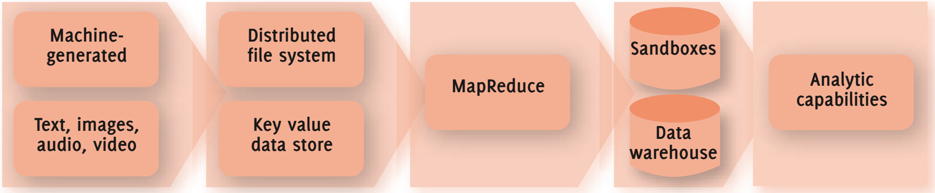
Structured



Step 2

We add unstructured (Big) data to this model:

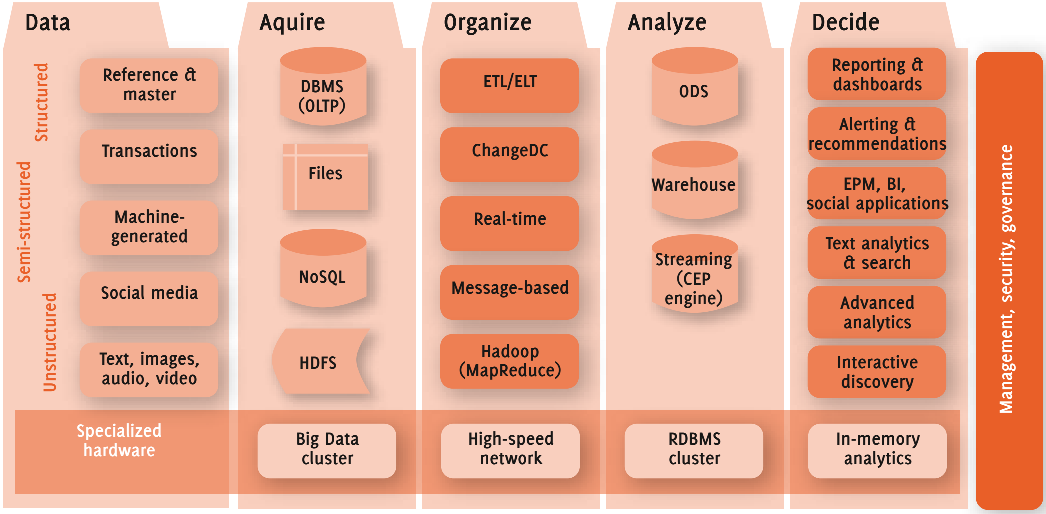
Unstructured



Step 3

The combination of 1 and 2 must be able to be appropriately scaled. Accordingly, a modern, integrated information architecture has a setup similar to the following, complete with specialized hardware, a high-speed network and in-memory analytics, among other features:

Integrated



Question 7: Which new technologies should I be watching?

Answer: the following three aspects of the technology behind Big Data are crucial points of concern for those involved in business IT: a) the three routes toward data integration (Hadoop, NoSQL and hybrid), b) ten Big Data core technologies for the entire process, ranging from data capturing to visualization, and c) the scalability of Big Data on the basis of the three Vs.

Routes to data integration

Some data integration solutions only work with the Hadoop platform, whereas others can cope with all kinds of Big Data stores and also offer analytics to boot. Organizations must be perfectly clear which data stores are important to them, and should align their data integration products and processes to these. Richard Daley, the co-founder of Pentaho, supplements this as follows, giving a good initial picture of what Big Data involves in technological terms. Daley distinguishes between Hadoop integration, the NoSQL route, and hybrid Big Data integration.

1 Hadoop integration

Many Big Data initiatives are based on Hadoop variants such as Apache Hadoop, Cloudera, MapR, Hortonworks and Amazon Elastic MapReduce. NoSQL databases are used to a lesser extent, as are high-performance relational analytical databases such as Greenplum, Infobright, Netezza, Teradata, Vectorwise and Vertica, among others. In a formal technical sense, Hadoop is actually a form of NoSQL.

Even the most basic Hadoop integration solutions go a step further than Hive integration these days. They support the entire range of Hadoop interfaces such as MapReduce and HDFS, and Hadoop eco-systems such as Pig, Sqoop and Oozie.

- Apache Hive is a powerful data warehouse application for Hadoop by means of which data can be accessed via Hive QL, which is a language similar to SQL.
- Apache MapReduce is the software framework for Hadoop, enabling the parallel processing of large quantities of data on large computer clusters.
- Apache Hadoop Distributed File System (HDFS) is the storage system for Hadoop applications. HDFS reproduces data and distributes the results to the computing nodes of a cluster for reliable and extremely high-speed processing. Some Hadoop

distributions offer alternatives to HDFS, such as NFS (from MapR) and Apache Cassandra (from DataStax). These give even better performance and can cope with data under Windows, OS X and Linux.
- Apache Pig is a script language for MapReduce programs under Hadoop. Due to their structure, Pig programs are easy to run in parallel, so that they can deal with extremely large datasets without difficulty.
- Apache Sqoop is a tool enabling the easy migration of bulk data between Apache Hadoop and structured datastores, such as relational databases. A simple command-line tool, Sqoop (SQL-to-Hadoop), positions tables or even complete databases in HDFS files, generates Java classes for processing, and makes it possible to place information from SQL databases directly into the Hive warehouse.
- Oozie is a server engine allowing Hadoop to run workflow jobs, such as MapReduce, Pig, Hive, Sqoop, HDFS operations and subworkflows.

2 The NoSQL route

If you opt for the NoSQL database route instead of Hadoop (although in technical terms Hadoop is also a form of NoSQL), your Big Data solution will have to support MongoDB, Cassandra and HBase. MongoDB is rapidly gaining in popularity and offers an easily accessible, scalable, high-performance, open-source NoSQL document store.

A NoSQL data-integration solution must provide the following facilities:

- *A user-friendly visual development environment* that every IT worker, data analyst and business user can deploy to manipulate, visualize and explore data and to report on his/her activities.
- *Hybrid data integration*, so that the NoSQL database can be directly used as a source for reports and dashboards. Other data must also be able to be added to a data warehouse to allow a 360-degree view of the business.
- *Integration of the NoSQL database* with existing Big Data and traditional data stores to give a complete data management and data analytics solution.

3 Big Data integration with Hadoop and NoSQL

Instead of focusing on data integration through Hadoop or NoSQL, you can combine both in order to link up to various Big Data sources and enterprise stores. This kind of overarching integration must perform four tasks:

a. alleviate the limitations of Big Data storage and processing, so that Hadoop is no longer subjected to delays when attempting to gain access to data in computer clusters, enabling unimpeded implementation of sort, group and join queries on NoSQL databases.

b. remove technical barriers. Users need simple, user-friendly visual development interfaces for high-performance data input, output and manipulation, regardless of whether they are working with the Hadoop or NoSQL Big Data platform. IT workers, developers, data scientists and business analysts must be able to tap into, integrate and analyze both Big Data and traditional data without undue difficulty.

c. guarantee integration with enterprise data. The Big Data platform must be integrated with enterprise datastores. Therefore there must be good tools to connect Hadoop and NoSQL databases to traditional relational databases, exchange formats, and enterprise applications.

d. provide a complete business-analytics solution with reports, dashboards, interactive visualization and exploration, and predictive analytics.

Ten Big Data core technologies

The processing stages that apply to Business Intelligence applications also apply to Big Data, but demand extra technological effort to enable the complete process of data capturing, storage, search, sharing, analytics and visualization to occur smoothly. In his book entitled *Big Data Glossary*, Pete Warden discusses a total of sixty technological innovations and provides the following concise overview of Big Data concepts and tools.

1 Data acquisition

(such as Google Refine, Needlebase, ScraperWiki, BloomReach, for example)
For accessing various data sources, internal or external, structured or unstructured. Most interesting public data sources are poorly structured, full of contamination and difficult to open.

2 Serialization

(such as JSON, BSON, Thrift, Avro, Google Protocol Buffers, for example)
At various points during processing, the data will be stored in files. These operations all require some sort of ranking.

3 Storage

(such as Amazon S3, Hadoop Distributed File System, for example)
Traditional file systems are unsuited to large-scale, distributed data processing, but nevertheless the Hadoop Distributed File System is actually suited to this function.

4 Cloud

(such as Amazon EC2, Windows Azure, Google App Engine, Amazon Elastic Beanstalk, Heroku, for example)
Hiring computers as virtual machines in a cloud environment is increasingly becom-

ing standard procedure. In this way, you can make use of large processing capacity for Big Data applications at relatively low cost.

5 NoSQL

(such as Apache Hadoop, Apache Cassandra, MongoDB, Apache CouchDB, Redis, BigTable, HBase, Hypertable, Voldemort, for example. See http://nosql-database.org for a complete list)
NoSQL (Not only SQL) is a broad class of management systems that deviates from the classical relational model.

6 MapReduce

(such as Hadoop and Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum, for example)
In traditional relational database environments, all processing takes place via a special query language after the structured information has been loaded. In contrast, MapReduce reads and writes unstructured data to all sorts of file formats. The interim results are passed on as files, and the processing is divided among many machines.

7 Processing

(such as R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, Bigsheets, Tinkerpop; start-ups: Continuuity, Wibidata, Platfora, for example)
Filtering concise, valuable information from an ocean of data is a challenge, but there are already many solutions that can help with such tasks.

8 Natural Language Processing

(such as Natural Language Toolkit, Apache OpenNLP, Boilerpipe, OpenCalais, for example)
Natural Language Processing extracts meaningful information from untidy, human-based language.

9 Machine Learning

(such as WEKA, Mahout, scikits.learn, Skytree, for example)
Machine Learning systems automate and optimize decision-making. The recommendations given by Amazon, for instance, are well-know applications of this function.

10 Visualization

(such as GraphViz, Protovis, Google Fusion Tables, Tableau Software, for example)

This is one of the best ways to extract significance from data. Thanks to interactive graphics, the presentation and exploration of information blend together.

Scalability of Big Data

In conclusion, let us examine how we can classify Big Data on the basis of the well-known basic trio of *Volume*, *Variety* and *Velocity*. To establish a picture of Volume and Velocity, it is interesting to check how long data transport actually takes from one location to another through a landline. Many people, used to normal domestic usage, have simply no idea of the time involved. Through a T-1 line, it takes 90 minutes to send 1 gigabyte from A to B. With Thin Ethernet it takes 14 minutes, and with Fast Ethernet 1 minute. This is not hyper-fast, but it seems reasonably acceptable. In contrast, 1 terabyte through a T-1 line takes almost 66 days. With Thin Ethernet this is more than 10 days, and with a Fast Ethernet connection it still takes 24.5 hours. Even with Gigabit Ethernet this still costs almost 2.5 hours. For high-speed Data Discovery iterations such as the 80,000 instances of the Capital One credit card company, this kind of bottleneck is unacceptable, and this applies to other scientific environments such as those of the Netherlands Forensic Institute or a genome institute. A thousand genomes amount to 200 terabytes, and we are still only speaking about the volume aspect of the matter.

In their overview article entitled "Tackling Big Data," Michael Cooper and Peter Mell of the IT Laboratory Big Data Working Group of NIST, the American National Institute for Standards and Technology, distinguish three Big Data types. They are related to the three Vs in the following way:

Volume	Velocity	Variety (semi-structured/ unstructured	Requires Horizontal Scalability	Relational Limitation	Big Data
No	No	No	No	No	No
No	No	Yes	No	Yes	Yes, Type 1
No	Yes	No	Yes	Maybe	Yes, Type 2
No	Yes	Yes	Yes	Yes	Yes, Type 3
Yes	No	No	Yes	Maybe	Yes, Type 2
Yes	No	Yes	Yes	Yes	Yes, Type 3
Yes	Yes	No	Yes	Maybe	Yes, Type 2
Yes	Yes	Yes	Yes	Yes	Yes, Type 3

Type 1 *Non-relational data representation required for effective analysis*
Type 2 *Horizontal scalability required for efficient processing*
Type 3 *Non-relational data representation processed with a horizontally scalable solution required for both effective analysis and efficient processing*

In other words: ***the data representation is not conducive to a relational algebraic analysis.***

Source: Cooper & Mell (2012)

Question 8: What is looming on the horizon?

Answer: With regard to your transformation and performance, you should aspire to belong to the category of the "Digirati," in line with the study published by MIT and Capgemini.

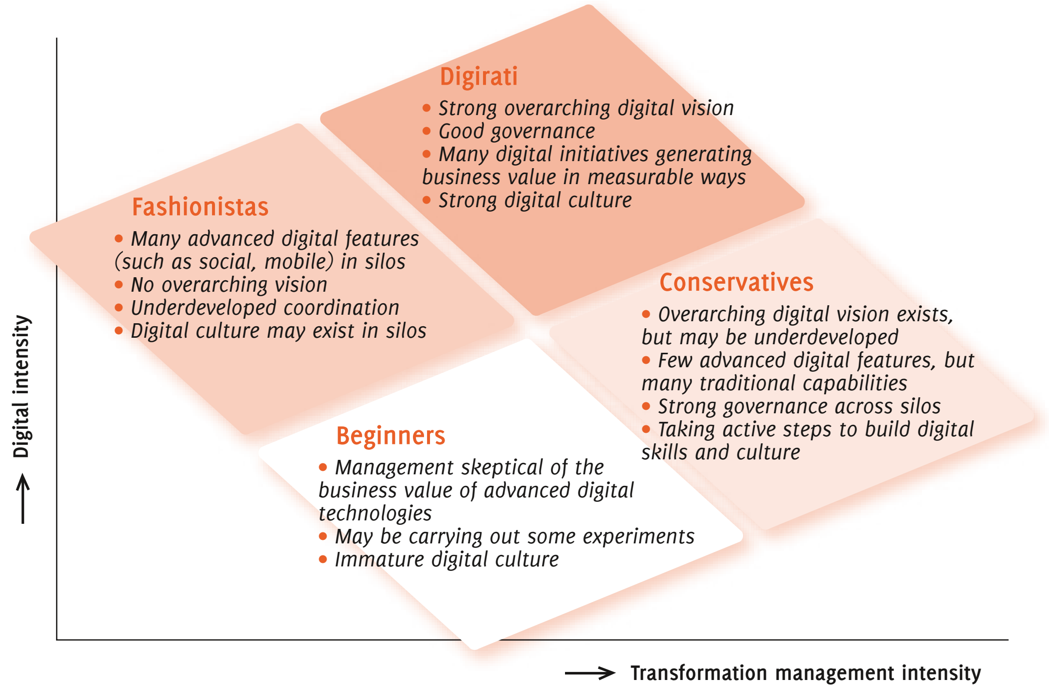
Big Data is a part of the digital path you are following. And there is good news about that path. As we mentioned in the Introduction, there is evidence that a digital strategy is rewarding in terms of turnover, margin and shareholder value. Such indicators are less relevant to the public sector, but here, too, it is worthwhile investing wholeheartedly in digital options. In May 2013, the Dutch government announced that all government services must be fully digitized by 2017. This should lead to annual savings of 300 to 400 million euros.

The potential returns of your digital activities, and thus now of Big Data in particular, is emerging in relevant studies. The results of recent research by the Massachusetts Institute of Technology (MIT) and Capgemini, as well as 391 other organizations, establishes a starting point for a longer-term strategy, a vision looming on the horizon of how you could evolve from a viewer into a top performer.

This study is called *The Digital Advantage: How Digital Leaders Outperform Their Peers in Every Industry* and it distinguishes four archetypes of digital intensity on the one hand and transformative capabilities on the other. The true "laggards" are politely referred to as "Beginners." They do not really believe in the power of digital technology, and perform a few half-hearted experiments at most. Their real problem is that their ambition to change is insufficiently developed.

This appraisal does not apply to the "Conservatives," the group that is depicted next to the "Beginners." Their transformative capabilities are actually high, and this is expressed in the way in which this group regards digital technology. The "Conservatives" do have a vision for digital technology, but they really ought to be more enthusiastic. In contrast, their business sections are well managed, and work is genuinely being done with the aim of realizing a better digital culture. However, there is almost no advanced IT in the departments.

This is the reason why we have depicted the *Digital Advantage* quadrant as a diamond here. The "Conservatives" are better off than the "Beginners," as the *Digital Advantage* study clearly shows.


Digirati
• Strong overarching digital vision
• Good governance
• Many digital initiatives generating business value in measurable ways
• Strong digital culture
Fashionistas
• Many advanced digital features (such as social, mobile) in silos
• No overarching vision
• Underdeveloped coordination
• Digital culture may exist in silos
Conservatives
• Overarching digital vision exists, but may be underdeveloped
• Few advanced digital features, but many traditional capabilities
• Strong governance across silos
• Taking active steps to build digital skills and culture
Beginners
• Management skeptical of the business value of advanced digital technologies
• May be carrying out some experiments
• Immature digital culture
Digital intensity
Transformation management intensity

The performance of organizations that exhibit high digital intensity and good transformative capabilities overshadows that of their peers. Top performers are the so-called "Digirati." They have a strong view of digital affairs and a digital culture. This type of organization is well managed and there are (therefore) many digital initiatives that palpably add value.

The remaining category, the so-called "Fashionistas," is carried along on the waves of fashion, as the name indicates. From the outside, this type of organization has a reasonably attractive digital profile, in the social media and in mobile applications, but the initiatives are fragmentary, a true overarching digital vision is lacking, and collaboration with the organization leaves much to be desired. This being the case, the "Fashionistas" may indeed be hip, but it could all be a good bit more efficient and sincere, certainly with regard to their transformative skills.

The interesting aspect of these kinds of archetype is that you will never actually find an actual organization that is identical in all aspects with one described in the model.

An archetype is a reference category. The significance of archetypes is that we are compelled to look closely and candidly in the mirror and investigate just who we are in certain respects, and how we can perhaps change for the better. This makes the *Digital Advantage* study a useful aid when formulating a plan to determine a strategy aimed at obtaining more benefit from Big Data. The study does not place the emphasis on Big Data, as already mentioned, but rather on solid data integration and the importance of a data strategy.

Looking to the horizon and transitional stages

Embracing "Digirati" characteristics can help you get your Big Data initiatives off the ground. That involves embedding Big Data in a greater body of digital strategy and strategic necessity. The study provides the following four transitional stages as intermediate targets:

1 Working on a transformative vision

Imagining how the organization might look in the future is a way to change the organizational mindset. On the one hand, it makes clear which rules will no longer be valid in the future and, on the other, it offers a glimpse of a new way of working. Enough books and aids (such as TheIdealFrame.com, for example) have been written about framing ideas, with the goal of providing a positive outcome. Greg Sattel gives an example of how to frame a business transformation:

> *"[...] business models can no longer be treated as stone tablets divined by wise men on mountains to last for eternity."*

And another one, formulated by Clay Shirky, is quoted by Kevin Kelly:

> *"Institutions will try to preserve the problem to which they are the solution."*

2 Digital governance is an important trump card

To make it more certain that investment in Big Data will be successful, it is necessary to examine the organizational governance structure, particularly with reference to next-generation Business Intelligence. The text at question 5 provides some reference points for this examination.

3 Members of staff must be involved and must remain so

Only by involving employees at an early stage can a Big Data integration strategy be truly successful. An internal innovation jam for ideation, implementation, and execution of Big Data innovations has proven to be a suitable instrument for this.

4 Collaboration between business and IT

Ensure that teams from business and IT work together to actively execute data integration. Intensive data integration requires the adaptation of architectures and processes. For example, marketing can direct its campaigns to smaller segments thanks to localized data. In hospitals, analyses of DNA deviations that initially took two years to perform can now be done in a few weeks. Methods of diagnosis and treatment are thus altered.

Question 9: What does this mean in organizational terms?

Answer: Actually making your organization data-driven begins with a good strategic plan to enable data, analytics, tools and people to jointly create business value. The directors, technology professionals, data scientists and managers must jointly determine where the greatest profits are to be obtained, and they must then select the first projects.

Let us take a concrete example. In his 2006 bestseller *Marketing Genius*, business guru Peter Fisk adeptly explained how most organizations deal with data. Fisk thought it foolish that everyone is constantly seeking more data. Most of the information that organizations accumulate, according to Fisk in 2006, is completely worthless. The real issues remain unaddressed as the data is aggregated and averaged much too quickly, so that all relevant information evaporates. It is often the case, Fisk argues, that management has already decided on its plans and it makes no difference whatsoever what the analyses show:

> *"Most organizations are weighed down by research reports, tracking data, analytical spreadsheets and the like. The research industry is consequently huge too, perpetuating the myth that more data is good for you. Yet most information collected by organizations is useless. It doesn't address the issues important to them, it is quickly aggregated and averaged so that any useful knowledge is smoothed out, and it is more often than not requested by managers who already decided what they want to do, irrelevant of the research findings."*

You certainly do not need to be a marketing genius to recognize the futility of the practice Fisk has identified. If this is to become Big Data practice, it should only be used to support decision-making in a formal sense, as it offers no solid advantage. This is grist to the mill of the cynics who claim that BI has not kept its promise and that Big Data is following in BI's slipstream. If new insights into structural innovation cannot get anywhere, continuing to collect data is no more than an empty ritual that should be halted. Fisk candidly commented that you should know what you are engaged in and which questions you want to have answered:

> *"The first steps in achieving insight are to stop [...] collecting more data than you need, resisting the desire to research everyone constantly, and to ask every possible question. There is also a temptation to jump into research without clarifying objectives, often finding that it has no particular purpose, or cannot answer the most important questions."*

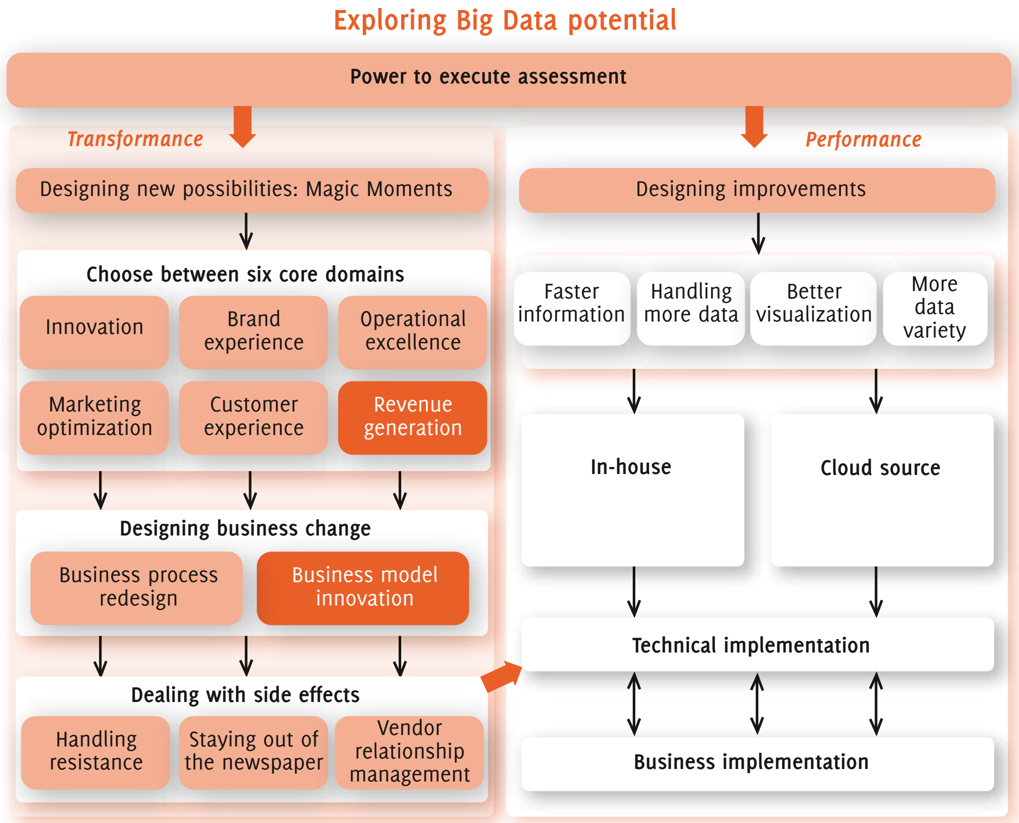
In his caricature, which undoubtedly bears a grain of truth and was intended to provide a wake-up call to organizations, the so-called "plan" that management cobbled together and intended to implement was not formulated in a methodical way. There is no real mention of a direction and route that deserves the epithet of "plan."

In March 2013, seven years after Fisk's warning and despite evidence that Big Data initiatives, if well applied, easily repay themselves, McKinsey expressed more or less the same sentiment. Most organizations had no clear plan to become truly data-driven, although those who do seriously work on this development soon achieve easily 5-to-6 percent more productivity and profit than the competition.

Making your organization data-driven begins with a good strategic plan of how data, analytics, tools and people can jointly create business value. This type of plan is, according to McKinsey, the common language of the directors, technology professionals, data scientists and managers when determining where the greatest profit can be achieved and when selecting the two or three data-intensive projects with which you wish to start.

With regard to Big Data, McKinsey draws a parallel with the situation forty years ago. Well-formulated strategic plans were more of an exception than the rule at that time. The pioneers booked striking results through these plans and an increasing number of companies slowly began to see the light.

Keeping this in mind, it is advisable to determine your execution capability, as in the drilldown that VINT proposes in the following figure, and allow yourself to be led by the lines of transformation and performance. Following these proposals, your organization will automatically come face to face with the most important issues that you will have to deal with.



This drilldown gives an overview of the broad contours of the organizational aspects. We wish to elucidate four particular aspects here.

Magic moments and practical improvements

The true "Digirati," the companies that have already had much more experience with crunching unstructured data, are primarily seeking improvement in performance: transporting data through a line from A to B more quickly, better visualization, integrating even more external and internal data, etc. These companies will also be more willing to look to cloud solutions. Organizations that are largely engaged in exploration begin by creating magic moments: a creative start-up stage in which ROI thinking is taboo for the present and long-protected customs are abandoned in order to arrive at genuinely transformative ideas.

Risks

In the lower left-hand corner of the figure you will see that there are side effects that you will have to take into account: the risk that you may be portrayed negatively in the news, for example. Whatever the case, *Privacy by Design* (see part IV) offers you the necessary guidance. However, resistance within the organization – because you are going to manage processes differently in the future, for example – is also a potential side effect. Vendor relationship management, the reverse of customer relationship management, is also a potential danger. If common citizens and consumers can gain more control over their own data, business models will arise in which these groups will ultimately – and more often – be able to claim their just rewards.

Strategic choices

In part I, entitled *Creating Clarity with Big Data*, we presented six general areas where organizations could readily begin with data integration and Big Data. These are:

- innovation
- customer experience
- brand health
- marketing optimization
- turnover generation
- operational efficiency.

It is self-evident that we should begin at the spot that is most important to the organization: at the heart of the business and in compliance with the strategy. To Amazon, for example, this is customer experience. On top of the mountain of data, the algorithm that also recommends books to you is an expression of this strategy. To Shell this "innovation" lies in the tracing and development of energy sources. The Netherlands Forensic Institute has operational efficiency as its the impetus behind its chain of investigation. A party such as Equens started up relatively recently and has chosen to use Big Data for turnover generation, by means of which it differentiates its earnings model. An exception may consist of initiating Big Data projects for general use, such as some telcos and organizations such as TomTom have done. The idea behind this course of action is to cultivate understanding for the Big Data activities of these companies.

Innovation at Shell

For a company such as Shell, innovation is an extremely important area where much effort and investment are devoted to Big Data. To an increasing extent, it is becoming difficult for Shell to access conventional sources of oil and gas. Seismic research produces 50 to 100 petabytes of data. In bygone days, a maximum of a few thousand sensors were deployed, but nowadays this may be up to a million. In a new unit, called "Technical and Competitive Information Technology" (Tacit), work is being carried out on new techniques for data visualization in 3-D and 4-D. This unit gives substance to the strategic intention to become the most innovative energy company. Shell foresees that it will need a hundred times more computer power than it has at present in order to reach its goals. There is a great deal to play for. A difference of thirty meters in a deep-sea drilling venture can determine whether or not oil may be found. Each drilling costs 100 million. In that case, it is advisable to have the very best analytics at your disposal, as well as the machines that can process and visualize the data. In Tacit, Shell collaborates with MIT, Stanford, Ferrari, the Chinese Academy of Sciences, HP and Intel, among others.

Operational efficiency at the NFI

The Netherlands Forensic Institute (NFI) has to deal with petabytes of data annually. The request it often receives is whether or not it can distill criminal behavior from a large variety of data within a relatively short time. This may range from a Wordfeud conversation to data on a damaged harddisk to data in a TOR network. In short, there is an enormous mountain containing all kinds of snippets of information, from all over the place. Time also plays a major role in the formula. In the meantime, the NFI has become so quick in tracing child pornography, for example, that the judicial system is now beginning to suffer from congestion. The speed with which the NFI can make a statement is often essential, for sometimes a suspect has to be released after a maximum of 72 hours in custody.

The amount of data and data sources to be examined in criminal cases, mostly in fraud, murder and child pornography cases, is increasing exponentially. To enhance the effectiveness of this investigation, the NFI has developed an advanced software application, Xiraf. This application can analyze large quantities of data at a high speed and make them searchable. A user, such as a police detective, can open the tool via a secure internet link and enter a search term.

There is also the Bomb Data System (BDS), a digital database that contains all the data on incidents and threats with explosives in the Netherlands. This state-of-the-art system was commissioned by the National Coordinator for Security and Counterterrorism. The BDS enables the storage of information on explosives and incidents with chemical, biological, radiological and nuclear substances in the Netherlands, and these can be directly digitally shared with other users. The BDS is also accessible via smartphone or tablet and therefore can be used at the scene of a crime. The speed of information processing is crucial in the detection process, in the field in the case of a direct threat, for example, but also in national trend and threats analysis and for forensic objectives.

New Equens data conversion cancelled

Until recently, Equens, the greatest pan-European processor of payment data, had never asked itself if it was possible to do more with the data at its disposal. After consultations with lawyers and discussions with the banks, Equens set up a service that intended to extract more value from its data. Encrypted and anonymized at card level, these data provide insight into spending at various identifiable locations. Real-time expenditure can be registered at shop level, in postal code zones and according to purchasing patterns. Simple questions that Equens could answer with a push of a button are, for example:

- What is the average expenditure of a consumer in a certain shop?
- In which other shops do my customers also spend their money?
- What were the expenditures in specified postal code zones yesterday?

These are the first steps toward innovating the earnings model. The approach is extremely cautious – *Privacy by Design* right down to the finest details – and ideas about new applications will not be hastily implemented but first thoroughly examined and discussed with experts and stakeholders. Nevertheless, the announcement of these plans did lead to some disquiet. The Minister of Finance even mingled in the debate and Equens announced that it would halt these activities for the present. Only if there is broad societal support will the theme be revived. Despite the fact that there seemed to be possibilities in a legal sense, the pressure to suspend the activities was overwhelming.

Question 10:
How does this affect everyday life?

Answer: In technological terms, questions 6 and 7 make the difference with traditional RDBMS environments, but a purposeful data focus on the combination of business, organization, and technology is both the core and the aim.

Conclusion and checklist

You are currently developing your Big Data potential of 2020 and, for this reason, this part concludes with a checklist of twenty questions. They have been taken from this book and from all interaction with the experienced experts with whom we have had contact online or in person. The issues of whether or not you are properly situated to tap into the Big Data potential, and of where the possible challenges may lie, can be answered using this checklist.

As a summary of this part and as an overture to the checklist, we first present the seven most important conclusions and recommendations in the mixture of business, organization and technology.

1 Structurally exploiting Big Data is becoming affordable

Only a few more years and the worldwide production of data will reach a stunning 40,000 exabytes, while the costs of working with that data will have diminished to 5 percent of current costs. Every sector and every organization can benefit from this.

2 Develop digital and organizational competences

Big Data potential lies in the meaningful combination of unstructured and structured data. Much work still has to be done on the digital side, but management must also do better. The Big Data potential of organizations lies at the interface of new digital and organizational competences.

3 Data Discovery is the next-generation BI

Insights from the new dataflows can cause a great deal of upset. This is one of the reasons why the concept of "digital transformation" is mentioned so frequently in relation to Big Data. Patterns become manifest and Data Discovery becomes a core component of the next generation of Business Intelligence.

4 Your aim is to belong to the Digirati category

Some organizations are better situated to realize their Big Data potential. They have been working toward a vision looming on the horizon for a much longer time, with the aim of making the organization fit for a digital future. We refer to them as "Digirati," and have given a description of their character traits in the section on question 8.

5 Ensure you have sufficient Big Data knowledge

Technological and analytical knowledge of processing unstructured data is a must in order to mix information from external and internal sources with structured data.

6 Big Data is "magic moments" and performance

Transformative Big Data initiatives begin with "magic moments": by choosing a domain in which your organization wishes to excel, while taking into account the risks and side effects. Performance Big Data initiatives are directed to existing projects with the aim of improving the performance. They involve the choice of what you yourself wish to do, and what you are going to (cl-)outsource (see question 9).

7 Big Data is an organization-wide evolution

In summary: you must develop a vision, acquire the right technological and data-scientific competences for the organization, give intensive data focus to the existing operation, and arrange your (data) governance properly.

In other words, you ought to read the following text attentively. We began with it, and it is appropriate to follow this part with it:

> Big Data is the new, intensive, organization-wide focus on business, organization and technology. Accordingly, you must make sure your organization has sufficient technological and analytical knowledge, as well as the appropriate digital and organizational competences. Your goal is to be able to excel in digital-operational terms. This is possible, because exploiting Big Data is becoming increasingly and rapidly affordable.
> With regard to Business Intelligence, *Data Discovery* is the next stage. It helps you combine lucid "magic moments" in your business operations with a significantly better performance via interactive visualization, exploration, planning and execution.
>
> To start with, you can stimulate your power of imagination in inspiration sessions, followed by one or more concretization workshops in which you determine, in conjunction with an organization-wide team, where lucrative Big Data initiatives can best be developed to suit your situation.

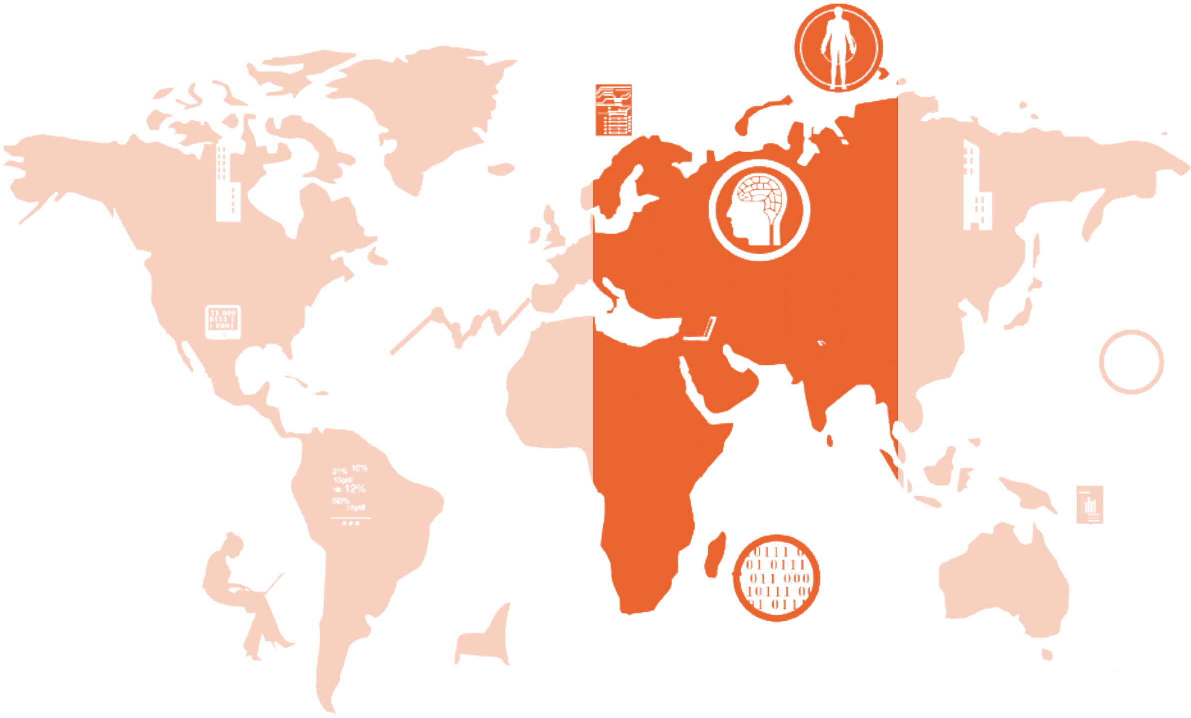
The following checklist further concretizes this intention and serves to structure your thoughts and plans and helps you formulate priorities geared to your own situation.

	Your Big Data potential	4	3	2	1
		very much			abso-lutely not
1	Does your organization have strong views on how Big Data could transform your business?				
2	Is the management of Big Data (governance) determined at the top?				
3	Are employees sufficiently involved in the Big Data vision?				
4	Is the collaboration between business and IT such that it can gain maximum advantage from Big Data projects?				
5	Do you have mainly Big Data projects in the areas in which the organization wishes to excel (strategic alignment)?				
6	Does the organization stimulate creativity in order to generate transformative Big Data ideas?				
7	Do you know where to recruit the scarce and elusive data scientists?				
8	If your knowledge workers were to receive the right data more rapidly, would that produce more profit?				
9	Do you possess data that may be interesting to others, and for which there is a market?				
10	Is it important to you to know about purchasing intentions at the earliest possible moment?				
11	Does data philanthropy (donating data to charitable projects) still give you favorable PR?				
12	Have you mastered the regulations for *Privacy by Design?*				
13	Is your Master Data Management in order?				
14	Do you know how to gain access to all your unstructured data?				
15	Have you prepared the infrastructure to cope with converting the further data explosion to concrete business value?				
16	Do you know which Big Data competences you are going to develop, and which you are going to tender out?				
17	Are you really capable of blending external data sources with your own data?				
18	Does your organization have the technology by means of which insights from data can be easily retrieved, instead of having to go searching for them?				
19	Does your organization possess a great quantity of data?				
20	Is there a great deal of diversity in your data?				

Part III

Big Social

Predicting Behavior with Big Data



1 What's next in Big Data?

Nine observations with both feet on the ground

This third part of *No More Secrets, Big Social*, offers a multi-faceted orientation to the promising Big Data development with regard to Social Analytics and Social Media. Some of these are indeed very promising, others only promise a lot. The more confident you are in your own judgment, the more able you will be to give a personal assessment and act in accordance with your observations.

The crucial question for now is simply *What's next in Big Data?* Many organizations are finding that they have been waiting too long for concrete solutions. These should be easy to implement, of course, and should amply repay the effort expended. Nervous eagerness and skepticism now mark the start of business practices that more and more will confidently build upon Big Data. Large Big Data projects and strategies are yet to come but the current emerging next practices that already can be discerned surely will find their place in daily operations everywhere.

In that context, please carefully consider the following observations, and see them primarily as central to the dynamic Big Data discussion that is currently in full swing:

1. Qualified best practices are under construction
2. Technology is making a breakthrough
3. Big Social will fulfill the promise of hypertargeting
4. Big Data runs the risk of becoming an out-of-control party
5. Clearly determine your actual needs
6. Behavior is often predictable without Big Data
7. Big Data builds upon traditional data centricity
8. Big Data ROI is simmering through
9. Social Analytics is the current Big Data pet topic

1 Qualified best practices are under construction

Anyone who has done any study on Big Data will surely know the promising report entitled *Big Data: The Next Frontier for Innovation, Competition, and Productivity* by the McKinsey Global Institute. It makes you think: so, something big is about to happen!

All the more remarkable, therefore, was the following admission by Michael Chui, one of the authors, at the MIT Sloan CIO Symposium in May 2012. Exactly one year after the publication of the *Next Frontier* report, Chui stated:

> *'There are no [Big Data] best practices.*
> *I'd say there are emerging next practices.'*

This seems to be at odds with the title of the report, but the coherence lies in the use of the word 'next'. Big Data will be capable of spawning benefits for innovation, competition and productivity, but convincing proof of this is not yet manifest. There is indeed a great deal about to happen, but much hard work must still be carried out in order to develop and implement 'emerging next practices'. With all the investments that have already been made, many organizations are not really waiting for such developments. Certainly not in the current economic malaise.

Organizations have begun to experiment, but it is as yet too early to give classic examples, best practices, that others can emulate. The Big Data domain is still undergoing too much change and it may even be likely that this will remain the case, in view of the claim of the favorable effect of Big Data on innovation and competitiveness. Both of course imply ongoing change and renewal.

2 Technology is making a breakthrough

In technological terms, a great deal is happening in the context of Big Data, such as the data-analysis language R. Another example is AMPlab at Berkeley University. With Big Data as its starting point, AMPlab orients itself to the combined forces of *Algorithms, Machines & People*. A Big Data milestone in the summer of 2012 was the appearance of the GraphChi software, developed at Carnegie Mellon University. This has enabled analyses on a common PC, where previously large computer clusters were occupied for hours performing such tasks. With a Twitter dataset from March 2010 as a benchmark, one single GraphChi PC turned out to be able to analyze this in 59 minutes. The previous occasion this was done, 1000 large computers spent 6.5 hours on the same task. The dataset in question is available free from the Infochimps.com website and contains 40 million users, more than 1.5 billion tweets, and 1.2 billion connections between users.

3 Big Social will fulfill the promise of hypertargeting

Regardless of how impressive this all may be, the big issue concerns the significance of it all. Not everyone is equally enthusiastic about Big Data; but many, including Jeff Dachis of the Dachis Group, hold the opinion that Big Data on social media forms the glorious hypertargeting future. Just think about it, says Dachis, hundreds of millions of people are busy on the social web,unguardedly sharing unconcernedly their whole lives with one another. It easily adds up to 500 billion dollars in brand engagement value.

At the beginning of 2012, Twitter had a total of 225 million accounts, and almost 200 million tweets were sent every day. In comparison: Facebook has more than 800 million active users and LinkedIn has more than 135 million. This 'consumerization' of Big Data will only assume larger proportions in the future.

There is skepticism about the use of advertisements on Facebook in particular. Just before Facebook was launched on the stock market, General Motors slashed its advertising budget for the social network. But this same GM still spends three times that sum on engagement with people on Facebook. The major challenge is to measure the ROI of this action. For such tasks we now have advanced Social Analytics tools and new algorithms such as GraphChi.

4 Big Data runs the risk of becoming an out-of-control party

One of the applications of Social Analytics is to gather as much information as possible on the online behavior of people – Big Social Data – with the aim of predicting what they are going to do next and what they are going to buy. Peter Fader, a marketing professor at Wharton Business School and co-director of the Wharton Customer Analytics Initiative, inserts a few prominent question marks here. He compares the projected goldmine of Big Social Data to Customer Relationship Management, which made its breakthrough in the early 1990s. At first it was regarded as the Holy Grail, but nowadays a harder evaluation is given: it causes huge frustration and is much too expensive; in short, the IT party has run out of control. Fader is afraid that things will turn out the same way with the current Big Data hype.

5 Clearly determine your actual needs

Dragging the entire Twitter or Facebook 'fire hose' through some Social Analytics refinery is simply nonsensical, says Fader. If you wish to get involved in hypertargeting, you have to look at tweets at individual level and link them to the transactions that a person executes. But online and mobile do not jointly form the complete new world that the Big Social Data evangelists, in particular, would have us believe. Of course, more information can lead to new insights, but the question remains as to how many data is needed for this? How interesting is it, actually, to know where someone is shopping at any given moment and what he/she is looking at? And which information on this subject should we retain?

6 Behavior is often predictable without Big Data

Fader believes that the real golden age of 'predictive behavior' occurred about fifty years ago. At that time, consumer information was very scarce. In the 1960s, Lester Wunderman began what he called 'direct marketing'. That was genuine 'data science'. Everything that could be known about a customer was kept up to date. What the direct marketing pioneers eventually achieved was RFM: the relationship between *Recency*, *Frequency* and *Monetary value*. The effect of F upon M is evident. R was the great surprise: it is easy to convince people to repeat previous behavior, even if they only buy things sporadically. However, you have to reach them immediately. In the marketing business, everyone is familiar with RFM, but it often signifies little to e-commerce people. With lots of Big Data you will undoubtedly come to the same conclusion, but that is a bit of a waste of all the time and effort expended. In this connection, a good eye-opener is the book *How to Measure Anything: Finding the Value of "Intangibles" in Business* by Douglas Hubbard, published in 2007. This is full of examples and tips to enable you to find out lots of things in a practical manner.

7 Big Data builds upon traditional data centricity

The majority of Big Data exercises take place mainly or even wholly in the existing data environment. If this environment is an advanced data warehouse with the corresponding tooling, substantial investment must have already been made here, large data sets will already be subject to examination, and organizations will not be particularly excited about investing money in Big Data and in solutions that are still currently under development, just for the sake of a few interesting experiments. Add the present-day economic predicament to this situation, and it will be apparent that the entire Big Data euphoria is not currently traveling under a lucky star. It cannot be denied that the data flow is increasing by leaps and bounds, but we have seen that coming for a long time now and it can be regarded as more of an evolutionary development.

8 Big Data ROI is simmering through

In the report *Big Data: The Next Frontier for Innovation, Competition, and Productivity* – which is still the directive publication *par excellence* – the McKinsey Global Institute indicates how easy or difficult it is to gather Big Data for each sector of the American economy, as well as what Big Data mining could contribute and what Big Data Maturity looks like, step by step. Despite partly fundamental reservations, this report nevertheless gives the impression that Big Data ROI can be gained within the foreseeable future. Eighteen months later, it still appears to be a little too early. It is even claimed that Big Data is not such an accurate term; Total Data Management or something similar might be better – is that what it's all about? The answer is: yes, this total data approach is essential and Big Data ROI is simmering through. We'll present some clear examples and directions in this part.

9 Social Analytics is the current Big Data pet topic

Social Analytics, the station between Web Analytics and the so-called Next-Generation Analytics, is a powerful antidote to all too grumpy Big Data skepticism. Gartner sees the latter as the immediate future. Here again we are confronted by that treacherous word 'next' and thus with the discussion about relevance with respect to Social Analytics, and of whether the glass is half-full or half-empty. Sullivan McIntyre of Radian6, a part of Salesforce, sticks to the first option and emphasizes the following: *'It becomes increasingly possible to make guesses about future behavior.'* Paul Barrett of Teradata also puts things in a wider perspective when he states: *'We are still in the early, black-and-white-TV stage of Social Analytics.'* In fact, both viewpoints can be said to hold true since sobering realism and cautious expectations go hand in hand.

Wanted! A multi-faceted orientation

The various discussion topics above and clear verdicts often being open require a multi-faceted orientation to be provided in this part. The main question now concerns the speed with which we will be able to switch from black-and-white to color, and then on to HD and 3D. In this part, which we have simply named *Big Social*, we wish to supply you with sufficient confidence to be able to draw your own conclusions and to keep a watchful eye on developments. We do this online at http://vint.sogeti.com/bigdata, where we are pleased to consult with you about the very latest perceptions of VINT and a select group of esteemed connoisseurs in the field of Big Data.

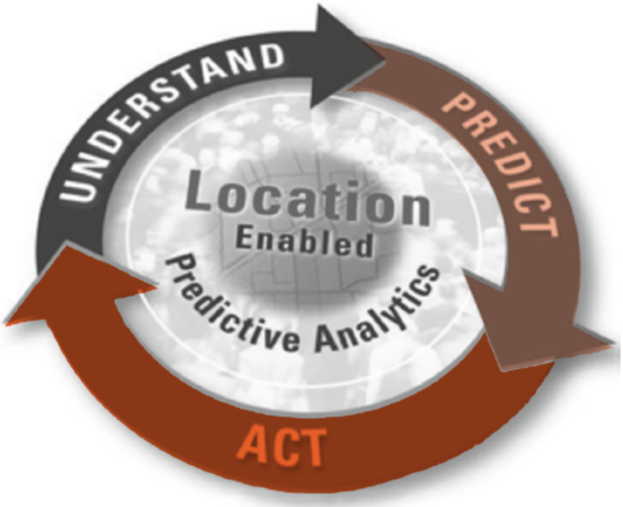
In the eleven remaining sections of this part we look at the many facets of what we concisely call *Big Social*. First, we'll examine the rhythms of human activity, and then respectively: data explosion's potential, the 'Big Five' social sources, the marriage of Web and Social Analytics, the development of Next-Generation Analytics, the shift to data and algorithms, the distortion of the social media lens, the Big Data technology toolbox, the need of listening attentively, the strength of Big Social Data. Finally, as part of the summary we will touch upon the organization of privacy, the theme of this part.

Although much more remains to be said, especially in these times of rapid Big Data development, these accounts and sketches are sufficient basis to form your own judgment, to draw conclusions, and to build innovative solutions in your own practice.

2 Rhythms of human activity

In June 2012, a remarkable 'emerging next Big Data practice' reached the news. Twenty-four hours in advance, it may already be clear where someone will be the next day, and sometimes the prediction has an accuracy of twenty meters (De Domenico, Lima and Musolesi, 2012). By involving the data of friends in the analysis of a separate people, a *Synchronized Rhythm of the City* can be ascertained. That insight lay dormant in a Big Data set, and originated from the smartphones of 200 participants. Without taking into account the behavior of friends, the accuracy of the location determination was only one kilometer. This improvement by a factor of 10 or 20 brought first prize to three British researchers in the Mobile Data Challenge competition run by the Nokia Research Center Lausanne.

To become familiar with human action, to predict something relatively simple such as the mobility patterns of people in a city for example, a great amount of personal data has to be processed. The iteration model needed for organizations to work on this comprises the following three-stage rocket, which, in this variant, is again related to the favorite social-mobile theme of location:



By examining an increasing amount of data from a growing number of sources, human actions are becoming increasingly predictable. The discipline concerned with this is often referred to as *Predictive Analytics*, and place and time are often favorite parameters, just as in the example presented above, where personal smartphone data was enriched with those of friends. In this case, the resulting algorithm produced an improvement that was at least ten times better. Although this is a research prototype, it is also a convincing demonstration of the power of more data from various sources in a real-time setting. In this form, that involves primarily *Variety* and *Velocity*.

For sure, there is no shortage of *Volume*. We leave behind digital traces, which often go further than place and time, wherever we go. Deliberately or unwittingly. When we go shopping, use a public transport pass, perform a search via Google, download an e-book, watch digital TV, listen to music via Spotify, are on the phone, send a mail, use a car navigation system, etcetera. Such data is available in abundance, to be read and utilized. Add the explosion of information that we share on social media such as Foursquare, Twitter, Zoover, Google+, Yelp, Pinterest, YouTube, Yammer and LinkedIn, and we have directly landed in a marketing heaven, where traditional focus-group research, surveys and sampling once constituted our limited resources in the framework of the trio Understand, Predict and Act. In our current Big Social era, that is now radically different.

What we want, from a marketing point of view, is to be able to zoom in and out on human behavior with an analysis instrument that is simultaneously a microscope and a telescope, and which also interprets a great deal of all the data too, so that we can make a timely and accurate offer. Or even better: can get paid immediately.

If a new product flops, we should not wait months to investigate the shopping sales figures. Signals from social media can give sufficient indication at an early date. This is called 'sentiment analysis', a concept that is well on its way to becoming the buzzword of the year. After all, it is via social media that we consistently ventilate our opinions about everything and more. It is the raw material of tomorrow's predictive mechanisms. That is the promise and the prospect: we will be able to predict who is going to commit fraud, where the following burglary will take place, what customers will look for next week.

Already there are many new things to be known, provided we tap other than our traditional data sources. Social is the aspect that brings great opportunity, and with Big Data and new analytical methods and techniques, we can make full use of the opening offered. Hypertargeting and personalization have never been within such short arm's reach.

3 More data for better answers

Imagine that you have several explanations for a fact you have at your disposal. But, if you now have to choose, what exactly are you going to select? It's ten to one on that you will say exactly what an aggravated Sherlock Holmes also said 120 years ago:

> *'Data, data, data! I can't make bricks without clay!'*

In our Big Data age, this fundamental response from *The Adventure of Copper Beeches* (1892) is more relevant than ever. We need more data, preferably as many as possible. Making a choice on the basis of a lack of information is certainly a rather anemic solution and, according to Big Data evangelists, it is totally irresponsible. Guessing is extremely dangerous, in the context of chasing the truth, and also businesswise. This is particularly relevant if your competitors do not have to rely on guesswork because they have taken the trouble to perform more profound search.

Mystery-solvers – whether they are called Sherlock Holmes, Columbo, or Wallander – never know where they should seek their data. Detectives grope in the dark and that is very frustrating, certainly if there is pressure of time. Businesswise, there is often no further option: a choice must be made quickly. Fortunately, the Internet has made things much easier in this respect: Big Social Data can guarantee increasingly better insight. That is the major difference and, accordingly, following one's intuition is becoming increasingly synonymous with the path of least resistance nowadays. Alleged talents, such as our gut feeling, can easier leave us in the lurch, but digital data will not, and they are there simply for the taking, in conjunction with the necessary analytical methods and techniques. We can place them under the denominator of *Social Analytics.*

There are roughly two Social Analytics schools. The first (Roebuck, 2011) is fully oriented toward social media due to:

- the growth of social media
- the growth of social media analytics tools
- the growth of social businesses on the basis of these two.

The other school, that actually entails the first one, sees Social Analytics in a broader perspective and orients itself to the analysis and prediction of human action on the basis of diverse sources. Mary Wallace, Social Analytics Strategist at IBM, counts herself as a member of this group. We will follow this broader viewpoint.

The following account is a classic example of how, in this case, a department store chain could quite easily acquire rather intimate facts from purchase data. The turnover of the American department store Target rose from 44 billion to 67 within a period of 10 years because people were able to segment customer groups well and to make them purposeful and relevant offers. This was largely due to the efforts of data expert Andrew Pole, whose work it is to examine great quantities of data until a hard predictive value has been reached. For example, Pole scrutinized the sales of baby products in relation to changes in purchasing habits over the previous months – such as a lotion with less aroma because women tend to develop a more acute sense of smell during pregnancy, food supplements in the form of zinc or folic acid, and disinfecting sprays for the hands. Pole thus identified 25 products by means of which he could even determine the date the baby was due.

NATE SILVER ON WHAT OBAMA SHOULD DO NEXT P.44
DON'T MOCK THE ARTISANAL PICKLE MAKERS P.14
A NANNY'S VIEW OF THE WORLD P.47
MANAGEMENT TIPS FROM DOWNTON ABBEY P.52
GREECE CONFRONTS ITS SPARTAN FUTURE P.18
The New York Times Magazine
HEY! YOU'RE HAVING A BABY!
How your shopping habits reveal even the most personal information. By Charles Duhigg

In Minneapolis, the local Target manager, who had no idea of how the offers were selected, could not explain to an angry father why his daughter was receiving discount vouchers for baby products. On closer inspection, it turned out that a data analysis of the store did get it right. This is all very smart and valuable for both parties, in principle, but customers must wish to participate and not feel that their privacy is being invaded. At the beginning of 2012, The New York Times Magazine published the much-discussed article on *How Companies Learn Your Secrets*, on the basis of this Target case and others. The cover immediately drew attention: *Hey! You're Having a Baby!*, which was constructed entirely of product packaging. The pay-off was: *How Your Shopping Habits Reveal Even the Most Personal Information.*

Hypertargeting as emerging Big Social trend

The Target practice is not just an interesting anecdote but rather an emerging Big Data trend. Big Social hypertargeting has even been industrialized by a company called MyBuys. It offers cross-channel personalization for online retailers and consumer brands. The company aims to drive engagement, conversions and increase revenue by capturing insights from individual behavior, then utilizing choice modeling algorithms to predict the products each consumer would most likely purchase. Underlying the MyBuys personalization engine is a Big Data repository of over 200 million consumer profiles and 100 terabytes of data, which the company uses to deliver real-time product recommendations.

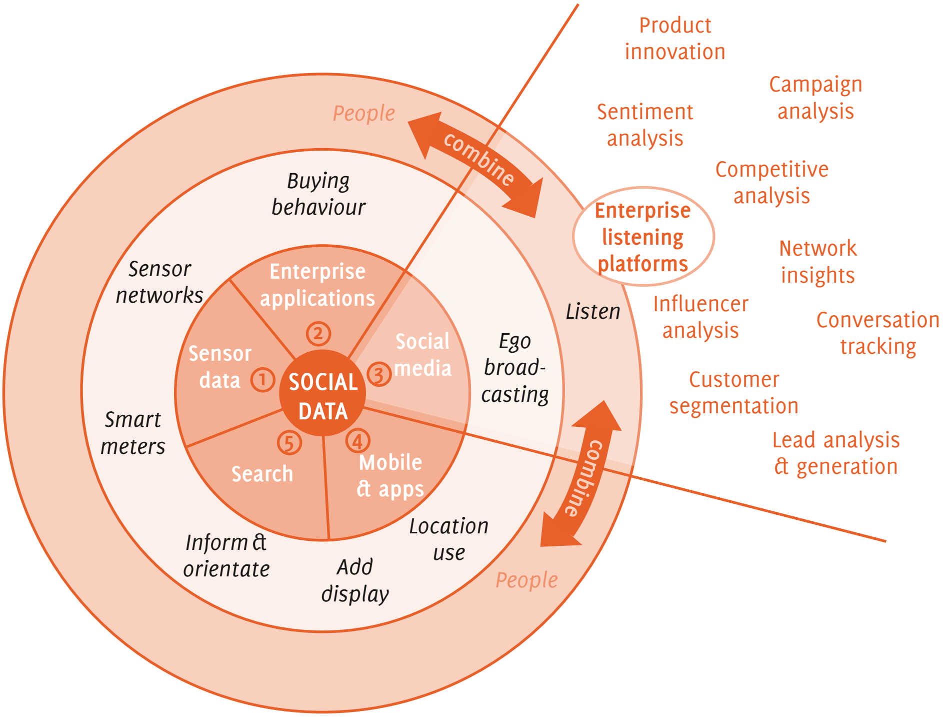
This Big Data development corresponds to organizational change. Just like the emerging roles of Data Scientist and Chief Analytical Officer, hypertargeting involving Big Data underpins the importance of the relatively new Chief Customer Officer role. According to the CCO Council, a Chief Customer Officer is *'an executive who provides the comprehensive and authoritative view of the customer and creates corporate and customer strategy at the highest levels of the company to maximize customer acquisition, retention, and profitability.'*

4 Total Data Management: the 'Big Five' social sources

Numerous data may form the basis of behavior analyses, such as client cards, search terms on internet, purchases, and responses to discount vouchers. In addition to the traditional enterprise applications as a data source, there are currently at least four other data categories that nourish the 'emerging next Big Data practices' of

Social Analytics in widest sense of the word 'social'. These are: mobile data and app data, search-engine data, sensor data, and semantic data (such as smart metering for example) and, of course, social media data.

Each of these 'Big Five' data sources has its own interesting characteristics. One is related to the way people perform searches on internet, another reveals the patterns behind purchasing. One type of data comes from one's own system, while another may come from an external system. Social Media Data concern the motives behind actions. We listen to these by means of 'listening platforms' and perform yet other analyses:



The 'Big Five' of Social Data, featuring the current Social Analytics practice for social media

It is essential to look at the whole picture and have a Total Data Management view, taking into account the strengths and weaknesses of data sets and the strength of smart combinations in the context of our three-stage rocket Understand, Predict & Act. But let us first examine each of the 'Big Five'.

1 Sensor data

These are data from (network) sensors, such as smart meters, which record the energy consumption and energy production of each household and neighborhood. The network consists of appliances that use energy, cars for the storage of energy in batteries, and people. This sort of 'neighborhood analytics' is a component of new production systems: a kind of 'Social ERP'. But other data, too, such as the tracking of purchasing behavior in shops ('in-store analytics' and 'anonymous analytics'), as well as numerous data from human-machine interaction also fit into this category.

2 Enterprise application data

Enterprise application data is traditionally used to recognize social patterns, such as purchasing behavior. It sits in structured databases of systems for Customer Relationship Management (CRM), Supply Chain Management (SCM), Enterprise Resource Management (ERP) or belongs to the data on a company's own website ('owned media'). In this context, we refer for example to on-site Web Analytics: the pages that people visit, the options that people click on, which 'landing page' is best in terms of leading to purchases, etcetera. There are also the so-called Cross Channel Attribution tools, which analyze great amounts of data from diverse sources (in-store, on-line and off-line sales). Enterprise application data are the building blocks of Business Intelligence, HRM applications, of production and commercial processes.

3 Social media data

This is all about data, often unstructured, that comes from individuals who are engaged in 'ego-broadcasting' on social media. This data is accessible to organizations. Corporate data from internal microblogs or company-based innovation platforms may also be an important source. So-called 'social listening tools' are applied in the analysis, enabling the subsequent steps of marketing and HRM.

4 Mobile data

This includes data from mobile applications, such as the popular apps category. Flurry and other players supply tools for App Analytics and make use of the same sort of metrics as those applied in web use. Mobile data may form the basis of location-based services that are supplied in real-time. Social media on mobile devices often also transmit location data. Social and mobile data together can be said to be a marriage made in marketing heaven.

5 Search data and external internet data (off-site)

Search data may come from search engines, such as Google Trends or Google Insights, from other suppliers that 'scrape' the search engines, from 'phoning home' software or ISPs. The data is used for trend analysis, Search Engine Optimization and Search Engine Marketing, by making use of keyword monitoring and services of Google Adwords, for example. Off-site web data offers better insight into the popularity of websites, into where the buzz is, or where comments are given. Such data may come from the logfiles of webservers or from page-tagging with Java scripts. This also is likely to occur of course on a company's own website: 'on-site internet data.'

Total Data Management

A Total Data Management view on the basis of these 'Big Five' is not only a vista on which organizations can work, it is also one of the prevailing trends we see among 'analysis vendors' such as Alteryx, comScore, Datasift, IBM, Infochimps and Salesforce. On Twitter, comScore characterizes itself vigorously as follows:

> *'comScore #measures the digital world. We manage #bigdata to bring you #mobile, #search, #video insights and more.'*

Data aggregator Alteryx appears to be quite exceptional:

> *'The only Business Intelligence company to offer packaged data from Tom Tom, Experian Marketing Services, Dun & Bradstreet and the 2010 US Census, and firmographics from the world's leading business data supplier in every license. This data, which spans spatial, demographics, household and firmographic market information, provides businesses with a deeper understanding of where, why and with who events occur. By analyzing this market data in combination with their own data, businesses are able to perform analysis that drives highly targeted and localized decision making, and improves the overall ROI of every piece of data.'*

Salesforce does something similar with Social CRM, by making combinations of social media data and CRM; or the Datasift company, which offers a cloud solution to enrich enterprise data with social media data.

IBM is currently working in Israel on an app to combine enterprise data such as historic purchase behavior and personal customer preferences: an interesting Big Data application for stores and supermarkets. The app places all data in an augmented-reality layer to guide customers through the shop via special personal triggers and offers.

> *'Beyond what IBM's augmented reality app may offer retailers on a customer-by-customer basis, it has the potential to create a treasure trove of metadata*

> *regarding product sales trends, hot in-store selling spots, traffic patterns and inventory issues, all of which could be used to maximize revenue per square foot, a critical metric in retail.'*

Google Now also mixes all kinds of social data, for individuals rather than companies, on the basis of our location, calendar, mail, searches, and video choices. What we need can be presented just-in-time, taking into account flight times and traffic jams.

At present, a convergence of data and tools is occurring all around us. This intertwining happened quite early, on the basis of our behavior on the Internet. For a good understanding of Social Analytics, it is important to know how this discipline evolved from Web Analytics. Then we will examine where it is going.

5 How Web and Social Analytics became entwined

Taking into consideration the behavior of customers, prospects and everyone who is involved in our brand and our products, attention is currently being devoted to the (Social) Web. In many cases, that is where most activity, representative activity or activity that generates most buzz occurs. Human behavior on the Internet has been the subject of analysis for a long time now. Early forms of Web Analytics were in existence as far back as twenty years ago. Social Analytics arose parallel to this, but much later, in 2006. In 2010, Web Analytics and Social Analytics embraced one another definitively.

In the framework of 'emerging next practices,' let us examine how that happened on the basis of a short timeline, spanning the last decade of the previous century and this first of this one. The data comes from the *History of Web and Social Analytics (1990-2010)* infographic by Webtrends and DK New Media. Accordingly, we first give two short descriptions of these two domains:

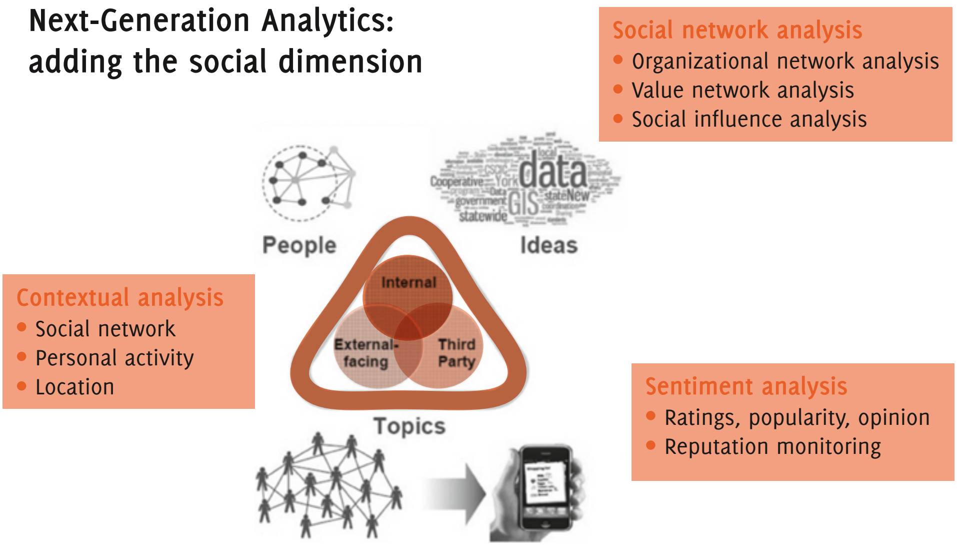
Web Analytics (1992–2010) according to Wikipedia:

> *'The measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage.'*

Social Analytics (2006–2010) according to Gartner:

> *'Social Analytics describes the process of measuring, analyzing and interpreting the results of interactions and associations among people, topics and ideas. [...] Social network [or media] analysis involves collecting data from multiple sources, identifying relationships, and evaluating the impact, quality or effectiveness of a relationship.'*

Social Analytics really arrived on the map in 2010, when Gartner included it in his top 10 of strategic technologies. Web and Social Analytics are now soulmates. The hectic dynamics depicted below in the illustration of *Web Analytics & Social Analytics 1990-2010* refer to companies that rise and merge (black), and 'emerging next practices' (red), which eventually all become mainstream. This process is ongoing, of course, and the interesting thing is that, in these times of Big Data, Web and Social Analytics are evolving to a further stage, which Gartner terms Next-Generation Analytics. In the context of 'emerging next practices', section 7 of this part looks toward a horizon beyond this concept of Next-Generation Analytics.



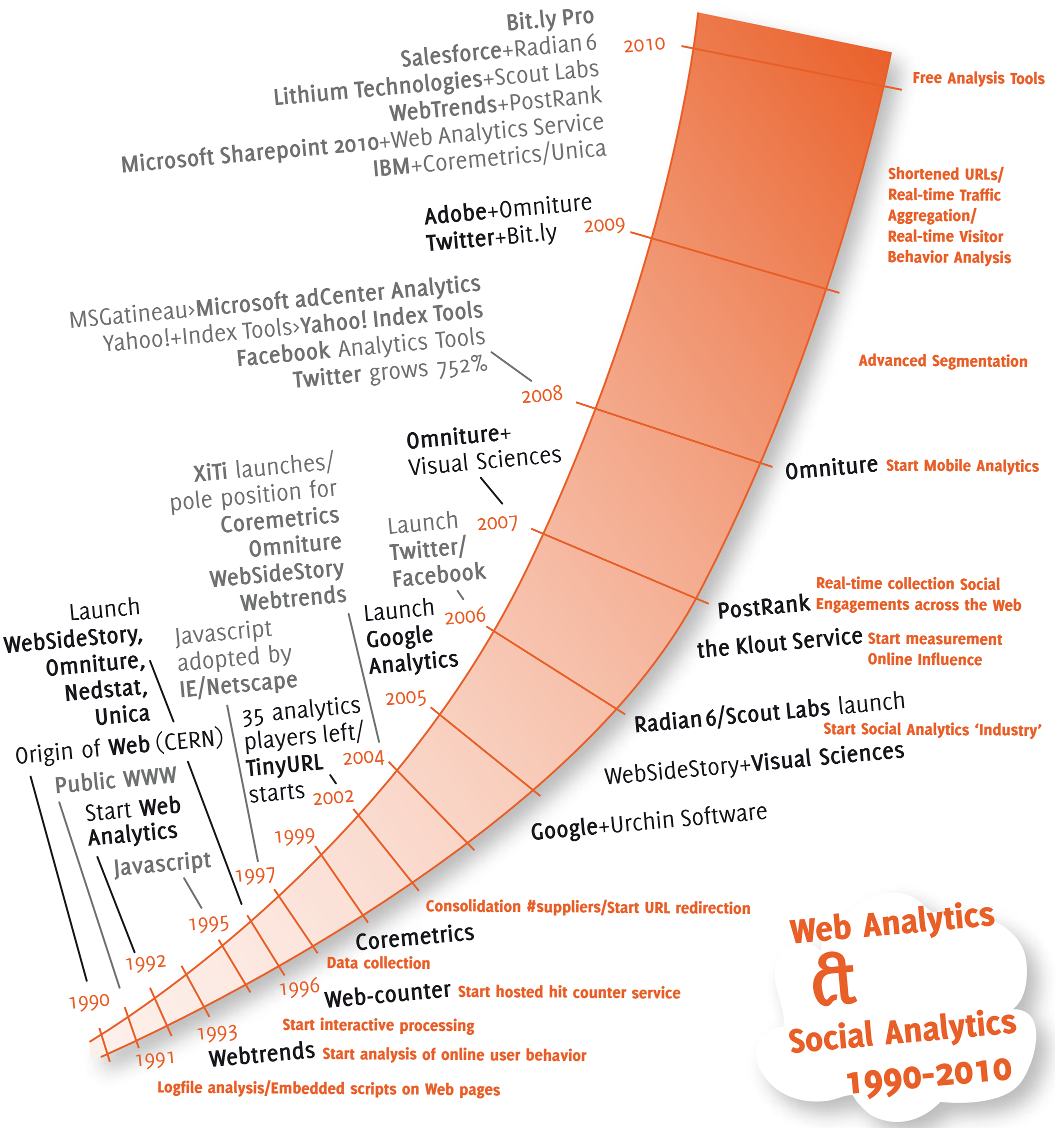
Next-Generation Analytics: adding the social dimension

Social network analysis
- Organizational network analysis
- Value network analysis
- Social influence analysis

Contextual analysis
- Social network
- Personal activity
- Location

Sentiment analysis
- Ratings, popularity, opinion
- Reputation monitoring

Source: Gartner, http://www.scribd.com/doc/81893261/February-9-Top-10-Strategic-Tech-Dcearley



6 Toward Next-Generation Analytics

Web Analytics and Social Analytics will be used together for some time to come, but the current focus is mainly on Social Analytics. If we were to pay tribute to someone who deserves praise in this context, that should be Lars-Henrik Schmidt. Since the early nineties, this Dane has been engaged in the development of a descriptive philosophical perspective, which he calls Social Analytics. Schmidt envisaged a discipline generally geared to reporting on the 'trends of these times'.

We now extract such trends from the trending topics of Twitter. Every minute we can check what people are communicating about, by counting how often a word with a so-called 'hashtag' (#) is tweeted. If we want to know the trends in different countries, we can use Twirus, for example. Social media enables direct up-to-date reporting. Enthusiasm about the use of social media tools is therefore very understandable. Analyses are available at the drop of a hat, there are free tools at our disposal so we can start immediately, and at the same time we can zoom out for the Big Picture and in on the (potential) buyers of products and services.

But Social Analytics – particularly Social Network or Media Analysis – is a transitional phenomenon, or at least, the excessive emphasis on this discipline is. The Gartner top 10 of promising trends for 2011 contained two datamining related domains: Social Analytics and Next-Generation Analytics. Of course we are moving toward this next generation: an integral Big Data approach to so-called Total Data Management, presented as a fundamental focus in part I *Creating clarity with Big Data*.

In the last section, we already saw Gartner's definition of Social Analytics. Next-Generation Analytics is typified thus:

> *'It is becoming possible to run simulations or models to predict the future outcome, rather than to simply provide backward-looking data about past interactions, and to do these predictions in real-time to support each individual business action.'*

How this 'Next Generation' progresses further in more detail is presented in the report entitled *A Framework for Social Analytics* by Susan Etlinger and Charlene Li, which was published by the Altimeter Group in August 2011. Here, too, the excessive emphasis on Social is merely a 'passing phase' so to speak, for *'Social is One of many Signals – Data are King'*, as we read in summary terms at the end of the report. Altimeter sees the future of Social Media Measurement as follows:

> *'"Social Analytics" or "social intelligence" will become an integral – and eventually indistinguishable – element of the enterprise's ability to sense, interpret, and recommend actions based on signals from the market. [...] One of the greatest impacts of the transition to the adaptive business is the advent of "Big Data"*

> *– the algorithmic increase in unstructured data that will stem from continuous interaction with customers, communities, and markets. […] Clearly this future state is several evolutionary steps away from social media monitoring and will require an entirely new processing and analytics approach that is able to make sense of both the unstructured nature and the sheer volume of data.'*

With this, Altimeter adopts the same position as Gartner with its division into Social Analytics and Next-Generation Analytics, namely: we are on the road from A to B, from Web Analytics and Social Analytics to *Total Data Analytics* as we like to call it, making the connection to our vision of *Total Data Management*.

Note: For a complete account see the section "Eight key Big Social definitions" at the end of this part.

7 Data and algorithms instead of models

The proper interpretation and linkage of data already leads to better decision-making, more sales, fewer risks and cost reduction. But, keeping in mind the 'emerging next Big Data practices' of Michael Chui, we can anticipate a great quantity of further improvements and possibilities:

> *'We are still in the early, black-and-white-TV stage of Social Analytics.'*

This is the view of Paul Barrett, Customer Management Director at Teradata. Let us examine what that literally means. Black-and-white was the TV period in which there were few channels and we required different antennas to receive them. Very often the only thing to see on television was 'snow'. We were troubled by 'atmospheric disturbance', and saw only 'snow' (or 'noise') if a strong wind had turned the aerial a little. It was far from being the ideal situation, with only gray tints to represent a colorful world. And, we could only see the same program at fixed times.

It would be an exaggeration to denunciate Big Social in a comparable way, because modern Social Analytics is in far better shape. But things could be better: a single antenna please, sharper picture, more details, more channels and sources, real time, various angles, more aggregation levels, pattern recognition and, above all: the ability to predict behavior. We need to know what people really want, serve them in a timely way, bind them to us, and build up relationships with and via them. This striking improvement is the commercial Big Data challenge for organizations.

If a dataset is large enough, and up to date, the empirical approach often works better than a formula. We could formulate a complex model to determine how any people will go down with flu, but investigating search results produces a faster and clearer picture. Gunther Eysenbach, a professor at the University of Toronto was the first to do so, in 2006. His conclusion was:

> *'The Internet has made measurable what was previously immeasurable: the distribution of health information in a population, tracking (in real time) health information trends over time, and identifying gaps between information supply and demand.'*

The same applies to many other things, such as the best pricing strategy for selling secondhand articles. We find that immediately in eBay data, which gives better insight into matters such as inflation and consumer confidence. In short, all kinds of answers are latent in large data sets, and we can uncover these without having to concern ourselves with models.

As far back as 2008, Chris Anderson of *Wired* magazine spoke provocatively about a Big Data vista, in which even theory-forming and the scientific method would become superfluous; but, of course, data cannot speak for itself. At most, empiricism and theory play leapfrog and, thanks to the data explosion, the emphasis currently lies on data and algorithms rather than on traditional models. This development has been ongoing for a number of years now; compare, for example, the statistical approach with that of machine learning:

> *'Statisticians emphasize probabilistic models for learning, and techniques for quantifying variation in the estimated model that results from variation in the learning sample. For many machine learners, the algorithm is the model, and emphasis is placed on developing interpretable yet flexible methods of learning in challenging context (computer vision, natural language).'*
>
> *http://www.crm.umontreal.ca/Machine06/index_e.html*

Scenarios on the horizon

In terms of predictive power and ambition, the latest developments based on Big Data and algorithms reach much further than traditional Web Analytics supplemented by dashboards to monitor Twitter and Facebook traffic and to give rapid response. Before dealing with this topic in the following sections, let us look over the horizon toward some interesting and realistic scenarios.

1 Drawing conclusions from apparently unrelated facts

We recognize this from online shopping, where we receive recommendations for items that we could be interested in. Sometimes the items are logically connected to our purchases: if we buy a digital camera, we are informed of an extra battery or memory card. But what if seemingly completely unrelated offers turn out be predictive in some way? For instance, does the speed at which you cycle past the bakery say anything about the chance of you going to the cinema this evening? Or is the moment of the day in which you play a game of Angry Birds indicative of the fact that you might be interested in a more expensive bottle of wine when you go shopping later? Or perhaps a 'like' on Facebook says something about your general health? Such links are latent in Big Data: patterns that have remained unrecognized until now. It is not inconceivable that systems themselves will go looking for correlations and that they will present options to us, proactively.

2 Creating lifelike personas

In marketing, and increasingly in IT – for they are inextricably linked in the context of Business Technology – we often work with characters: concrete personal descriptions that are characteristic of typical customers. At present, it is still a human task to develop personas on the basis of research and insight: who they are, what they are called, what jobs they have, which preferences, etcetera. By devising scenarios for personas, valuable improvements that are good for the customer, company, process, chain etcetera can be recognized. By combining the patterns behind all the customers, better, richer and more meaningful personas can be developed.

3 Predicting on the basis of 'live' behavior

Much analysis is still oriented toward facts from the past, combined with actions that are taking place at this moment. But what if the real-time component is applied more intensively? What if my telephone passes on exact data about what I think, feel and want, on the basis of my location, my actions in the previous twenty minutes, the pages that I looked up on the Internet, the apps I used, the sounds from my surroundings, my agenda, and the activities of my Facebook friends in order to determine my state of mind on the fly. Which themes are buzzing around in my head, what am I feeling, which urgent needs do I have, consciously or unconsciously? Google Now is an interesting impulse toward a working version of this facility, and the influence of technology such as Google Glass, which aims at continuous presence, could signify a giant step further in this direction.

4 *From predicting to subtly influencing*

When we have come so far that we can estimate, with any degree of certainty, someone's current frame of mind, the following possibility immediately arises: which minor and major impulses can we give someone to ensure that he or she enters a particular mental state, one in which he or she is quite happy, is open to experiment, and ready to dispense cash? Perhaps this involves music from a mobile telephone or activating a picture on Facebook, in order to set the right tone. Perhaps the route needs to be adjusted a little so that we ride through a tree-lined avenue? If enough people participate, a subtle variation will result in optimal influence. This may have a beneficial effect: to help people lead a simple, valuable and happy life; but the possible darker side of misuse and intrusive manipulation is no less realistic.

5 *Really smart organizations*

Nowadays, everything in an organization is digital: e-mail, telephone, content, finances, access doors, light, climate, presentations, training courses, you name it. Can we discover more patterns here? Perhaps only in the larger organizations initially, where the quantity of data is big enough to gain significant insight, but eventually it will be available as a service to every company, for instance via the Big Data algorithm set of providers like MyBuys. It will enable benchmarking in relation to successful companies, an optimization of processes on the basis of the way in which work is currently executed, an understanding of problems and opportunities, policies to cope with risks, and perhaps even proposals for the more human aspects, such as training courses, vacancies or evaluations.

6 *Validation by means of variation*

At present, Big Data is still primarily a matter of drawing upon data flows that already exist and attempting to formulate conclusions. But what would happen if the system itself were to go searching for new data, if it could try out things itself, by means of intelligent machine learning, in order to see what the effects are? For example, it could send a news bulletin to certain people to see if they do anything with it? Or perhaps send an SMS with a fact from all Big Data to examine whether or not a threatening transgression of certain measurements can be avoided? This is already very commonplace in the world of web advertisements: the division of all visitors into groups, each of which receives a slightly different variant of an advertisement, even placed in a slightly different position on the page. By measuring which configuration has the most effect, the placement of advertisements can become increasingly efficient over time. If Big Social systems can undertake autonomous actions, variation can be embedded for the purpose of seeking and achieving maximum impact.

7 The end of unpredictability

This is most interesting, at least as a concept: if society, trade & industry and government authorities are all convinced of the importance of data, of the importance of searching for patterns and of better predictions about all kinds of topics, can we create a situation in which we can look forward, six months ahead for example, with reasonable certainty? Can we then anticipate movements on the markets, in stock markets and innovations, for example? Much of the present economy exists thanks to unpredictability: someone who is prepared to take risks in exchange for payment. If this risk diminishes in the near future because we know how people will react to something, what the risks are, and what the chance is of something happening, what will then form the core of our economy? As long as we cannot predict the weather, any prediction of society is probably out of our grasp. Although – how predictable are people in fact? Dirk Helbing and his colleagues have received 1 billion euros from the European Union for their *Living Earth Simulator* or *Future ICT Knowledge Accelerator and Crisis Relief System*. The name says it all!

8 The transparent human being

Much interesting research is taking place on the relationship between the subconscious and the conscious human brain. There seems to be consensus about the idea that the subconscious mind is mainly responsible for our behavior and that, to put it simply, identity and the 'ego' are spectators rather than helmsmen. It remains difficult to assess our own behavior. We may think, for example, that we find it important to live healthily, but many actions nevertheless indicate that health does not enjoy genuinely high priority. Reality is much more complex, of course, with interesting feedback between behavior, result and 'ego', but it is plainly evident that concrete behavior says more about who we are than about what we think. Big Social can play a useful role in giving people more insight: who are you, what kind of work best suits your qualities and which relationships have the best chance of success? Hard data as a crowbar to finally break open the 'conditio humana.'

9 Hackers galore

No system whatsoever can ever be completely safe. The technology around e-mail spam and cyber attacks has demonstrated that people with less than ideal intentions also make use of IT for their own profit. Big Social tools can also give valuable help to criminals. All kind of private data can be used to guess passwords. Or think about an automatic system that can simulate a good friend in order to separate people from their money or to acquire compromising data. In view of the amount of spam in the world, it is clear that people no longer shrink from causing enormous nuisance to gain benefit. So what would a criminal do with a perfect social database?

8 Social media as lens distortion

Following the metaphor of zooming in and out with various kinds of data lenses, any distortion of the lens can be of major importance. Before discussing the tools that are intended to improve your vision and expand your horizons, we shall first draw your attention to several notorious distortions.

1 Not all Facebook accounts are real

Marketers are crazy about Facebook 'likes', even if it only to corroborate the success of their own actions. Such likes should, of course, come from real flesh-and-blood clients. The BBC study entitled *Who likes my virtual bagels* illustrates that an absurd product can be successful on Facebook thanks to the likes from fake accounts in Egypt. Almost 9 per cent of accounts are fake, which Facebook itself admits. So that amounts to around 83 million accounts. In 4.8 per cent of the cases it involves double accounts; 2.4 per cent do not belong to real people – the aim of Facebook – but to an organization or a domestic pet; and 1.5 per cent of the Facebook profiles are used for spam, for instance.

2 Twitter is also a distorted lens

Almost half of all Twitter accounts are allegedly fake also. There is a well-known joke: if all humans perished tomorrow but computers kept on running, trending Twitter topics would continue for years. There are companies that do separate the wheat from the chaff and maintain databases with only real live twitterians and tweeps.

3 Do you understand what is being said?

Automatic text analysis is not perfect. The quality of Google Translate amply demonstrates this. Tweets and comments may be ironic or may be full of slang. So how can we determine what the message is genuinely about and whether it is positive or negative? Nevertheless, technology is also advancing in this field and some market players actually claim that their computers and software really do understand what is being said.

4 What exactly are you measuring?

If only it were so simple, that the success of advertising could be predicted on the basis of what people tweet or post on Facebook. Much research has been performed on the effect of advertising, providing a completely opposing picture. For example, the *Remember the ad, forget about the product* effect: humorous adverts score highly but sales do not always rise. And, with reference to washing powders, we know that they do not score highly on likeability but none the less can be very effective. Thus, the number of likes on Facebook is not a reliable standard for judging whether or not to continue with a certain campaign.

5 How reliable is ego-broadcasting?

Most biographers warn about autobiographies. Probably rightly, because we tend to enhance our own performance in self-written histories. We already broached that subject in our book *Me the Media* (VINT, 2007) – particularly regarding all those 'hyper-egos' that display a positive correlation with Narcissus in Greek mythology. What does that say about social media? How authentic or 'true' are those messages? Perhaps the unreal world will be experienced as the real world in the near future, a 'hyper-reality' as Umberto Eco among others calls it, a Disneyland 'that can give us more reality than nature can.' But even without these philosophical thoughts it is good to realize that data from social media must often be taken with a pinch of salt.

6 Can you hear what is not being said?

The situation outlined above provides enough reason to listen to what is not being said. Sherlock Holmes once solved a murder because a dog did not bark. All the things that are not tweeted, facebooked and pinterested can also be regarded as important information. Even if it is only because we know that some things are not readily included in autobiographies – despite all the openness and transparency on the Internet.

9 The toolbox is bursting at the seams

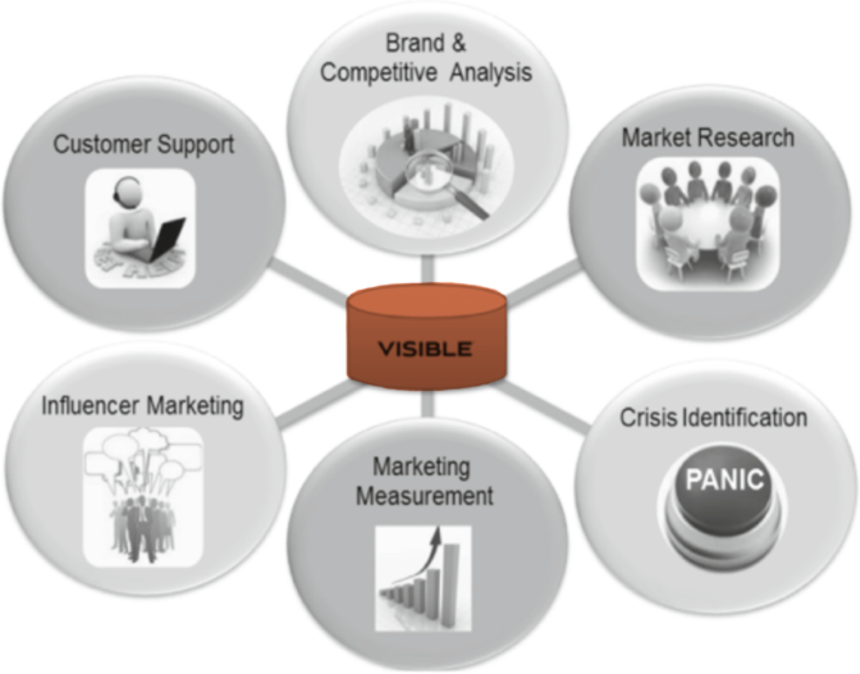
Despite and, of course, thanks to the above-mentioned critical remarks about predicting on the basis of online comments, the Social Analytics toolbox for companies has grown exponentially in the past few years. The following is only a small selection:

- Insight by Adobe
- Biz360 by Attensity
- Networked Insights
- Visible Technologies
- Scoutlabs by Lithium
- Radian6 by Salesforce
- Cognos Consumer Insights by IBM
- The Azure/Hadoop solution by Microsoft

- A whole range of tools for Twitter and Facebook, such as Osfoora, Tweetdeck, HootSuite, MentionMap
- The mobile-analytic tools by Google and Flurry.

For a fairly current overview of diverse so-called 'listening platforms' please visit Wikispaces (http://socialmedia-listening.wikispaces.com/Tools) or the Social Media Monitoring Wiki by Ken Burbary, co-author of the book *Digital Marketing Analytics*, which will be issued by Pearson Higher Education Publishers in early 2013. In July 2012, Ideya Business Marketing & Consultancy counted the number of resources available and registered almost 250 Social Analytics tools, of which around 50 are free. (Read a part of the report at http://ideya.eu.com/publications.html.)

Some suppliers bundle a great number of Social Analytics application areas to form a *suite* or a *hub*. To give an idea of important top domains, we now present the suite components of Visible.



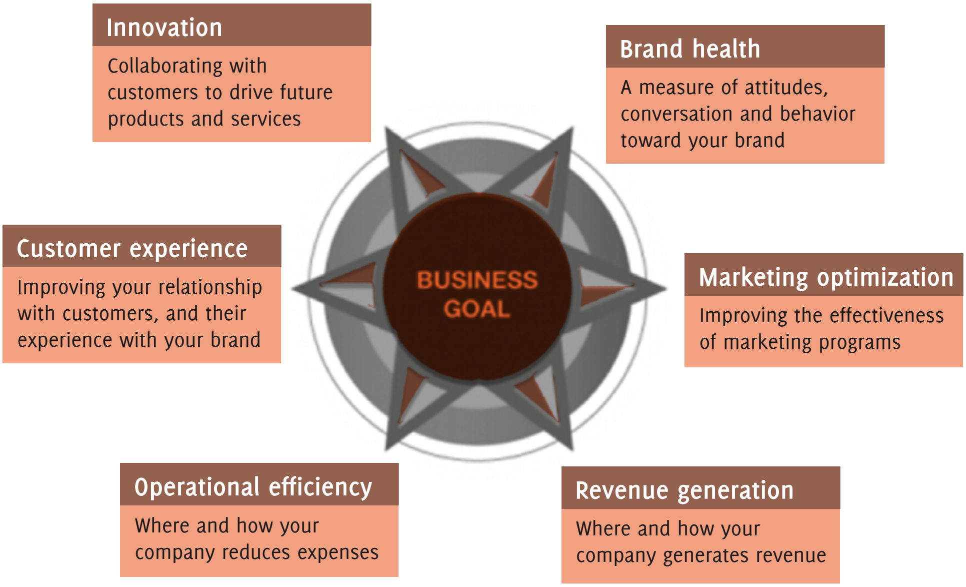
The Social Media Hub by Visible

These tools aim to enable us to delve deeper into the thoughts, intentions and behavior of people, with the goal of stealing a march on our rivals.

Social Analytics can be applied in a range of activities, from pre-sales, such as *lead generation*, to after-sales and *customer support*. With Social Analytics, we can assess the sales apparatus: identify and reward important people in the organization on the basis of the *influencer scores* from the chatter microblog on Salesforce for example. We can also perform analyses on a higher level of abstraction, and measure *sentiment* about certain customer experiences. This is all aimed at inspiring the marketing apparatus or product development. We measure the 'rumor around the brand,' 'buzz,' 'word of mouth,' or make comparisons between our brand and that of our competitors. We segment markets, divide customers into types (personas) or into locations whose coordinates are automatically sent via mobile apps. We can measure *whether or not an advertising campaign has been successful*, regardless of whether it has been geared to viral and social-media campaigns or to traditional TV advertising and published advertisements. We can examine who has which influence on social media and what these people can mean to our brand. And so on.

All such areas have their own terminology and acronyms. We are acquainted with the expression *Brand Protection*, and *Influencer Marketing* is the term used for spotting and influencing important people. This is all secondary to the possibilities that arose in the Google era (SEO, link-building, etcetera) and the enterprise application age of CRM, ERP and SCM. There are abundant possibilities, but that is not strange in view of the fact that *Social* represents life itself. The analysis of behavior is interesting for numerous corporate applications and that is why tooling is expanding in all directions. The acquisition of such tooling begins with the question about what exactly we wish to achieve with the tooling.

Altimeter has developed a framework for Social Media Analytics, which defines six rather obvious domains to link business sectors such as marketing, innovation and operations to targets. It is, indeed, the old familiar *Management by Objectives*. In the report we see that the focus turns successively to listening, acquiring insight, determining metrics and, of course, taking action. In the various domains, predictive models form the basis of the activities to be developed.



Innovation
Collaborating with customers to drive future products and services

Brand health
A measure of attitudes, conversation and behavior toward your brand

Customer experience
Improving your relationship with customers, and their experience with your brand

Marketing optimization
Improving the effectiveness of marketing programs

Operational efficiency
Where and how your company reduces expenses

Revenue generation
Where and how your company generates revenue

10 Start by listening attentively

Sullivan McIntyre of Radian6 states that social media data play an important role in the step from reactive analysis to predictive analysis. If you entwine your social media with other systems, *'it becomes increasingly possible to make guesses about future behavior'*. He presents three criteria that these data must meet:

1 Are the data real-time?

On line, social data moves fast as lightning. The freshness of the data is crucial.

2 Are there interesting metadata?

If we receive thousands of posts and have to analyze everything manually, the moment has soon passed. A rich set of metadata enables us to respond to trends quickly

3 Are the data integrated?

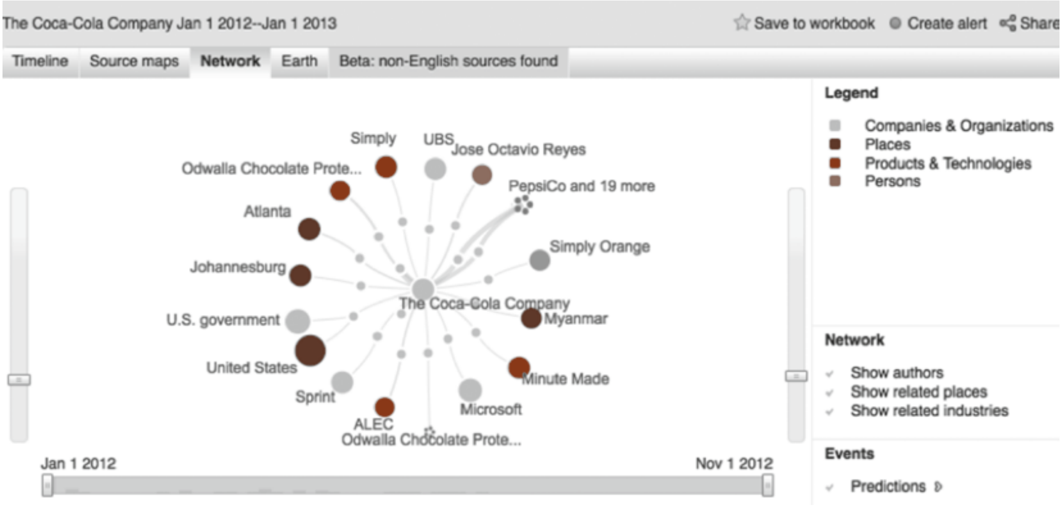
Data must be able to be linked, as much as possible, to other relevant sources in order to be able to undertake the appropriate action.

To be able to apply Social Media Analytics, we must listen attentively to what is being said. We do so by means of *Enterprise Listening Platforms*. Our three-stage *Under-*

stand-Predict-Act rocket presupposes that information is sent, received and interpreted. In this process, the influencer scores are key, as is a profile of the people being followed, a control center to keep the data up to date, a network analysis to determine connections, and sentiment analysis.

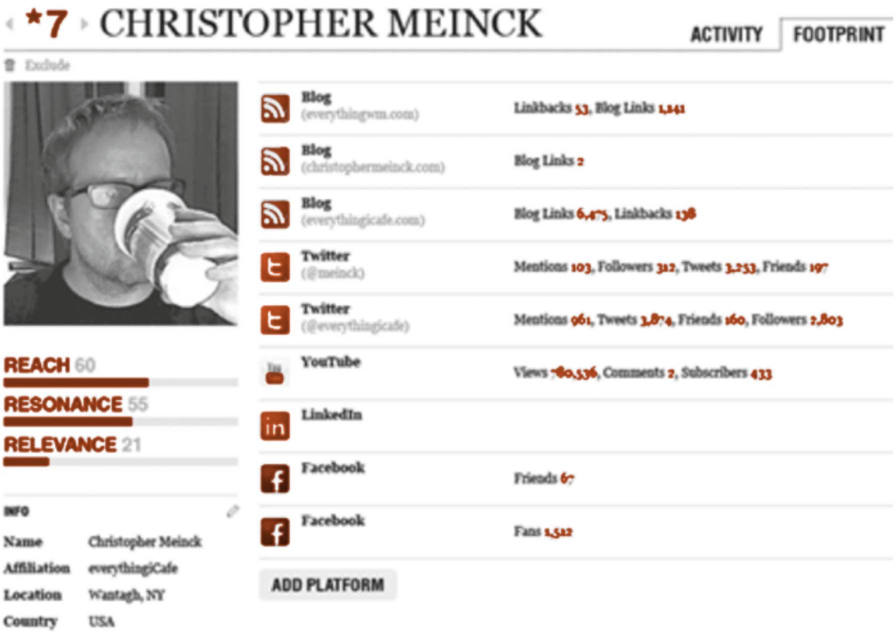
Listening platforms

Recorded Future, a company financed by Google and the FBI, is the name of a listening platform with a very specific mission namely to *Unlock the Predictive Power of the Web*. The company registers events that have not yet taken place. The organization uses blogs, websites and social media as input. It searches on 'next week' and 'next month', and delivers answers to questions such as 'Where is Obama going in August?' or 'Which beer festivals are held in October?' The launch of new products and technologies by competitors can also be monitored in this way, or perhaps conversations about a brand such as Coca Cola.



Influencer scores

There are also platforms that record the influence of people on social media; after all, we must maintain good relations with influencers. Twenty Feet has developed such a tool and calls it an 'ego-tracking service'. Ultimately someone is allocated a score, by the social network Klout, for example. Reports of influencers can be pretty detailed, such as this snapshot profile of Christopher Meinck made with Traackr.



*7 CHRISTOPHER MEINCK

REACH 60

RESONANCE 55

RELEVANCE 21

INFO

Name	Christopher Meinck
Affiliation	everythingiCafe
Location	Wantagh, NY
Country	USA

Blog (everythingwm.com)	Linkbacks 53, Blog Links 1,141
Blog (christophermeinck.com)	Blog Links 2
Blog (everythingicafe.com)	Blog Links 6,475, Linkbacks 138
Twitter (@meinck)	Mentions 103, Followers 312, Tweets 3,253, Friends 197
Twitter (@everythingicafe)	Mentions 961, Tweets 3,874, Friends 160, Followers 2,803
YouTube	Views 780,536, Comments 2, Subscribers 433
LinkedIn	
Facebook	Friends 67
Facebook	Fans 1,512

Profiling

But much more can be known than merely whether or not someone is an influencer. Male or female, car driver or not, country or residence, hobbies, married, divorced, etcetera. This kind of information is stored on the servers of IBM, which has gathered all the Twitter accounts in the world and removed the 'bots' from them. A profile has been generated on the basis of the content of the tweets and the person's own profile. At the moment that, for instance, a media company wishes to know if a new film trailer will be enthusiastically received by a certain public, this type of Twitter profile can have a predictive effect. A few minutes after showing their trailer, an American media company was already informed that the intended target group was not enthusiastic, and could also discern those who were more positive.

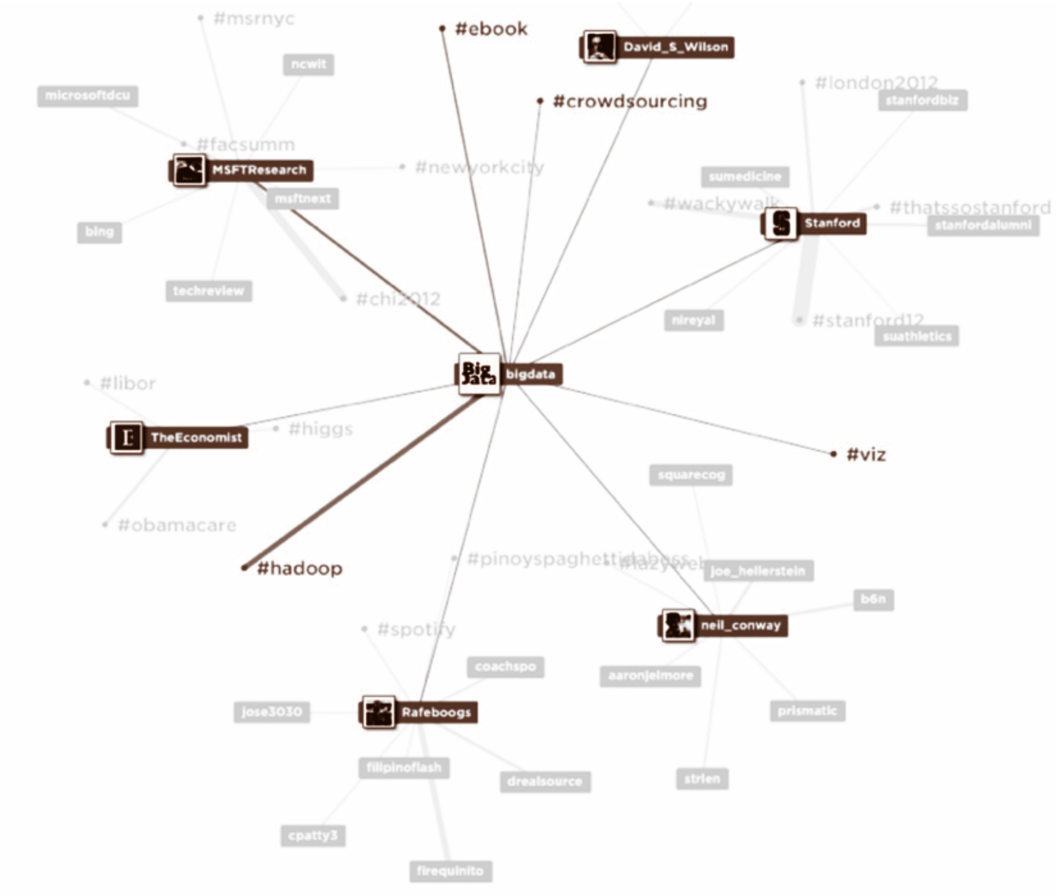
The Social Media Listening Center

The picture below is one of a Salesforce Radian6 dashboard on various screens in a Social Media Listening Center. To consumers, it is very irritating when a certain service freezes or breaks down. Via their listening platforms, the webcare teams of organizations record such messages and undertake action on this basis.

KLM is one of the companies that work in this way. The next time that someone tweets, the previous conversation can also be directly accessed. These dashboards also provide overviews by region. People follow messages on a world atlas via language and country. In this way, sentiments about the last TV ad or about the treatment at the reception desk are presented on the dashboard in the control center.

Network analysis

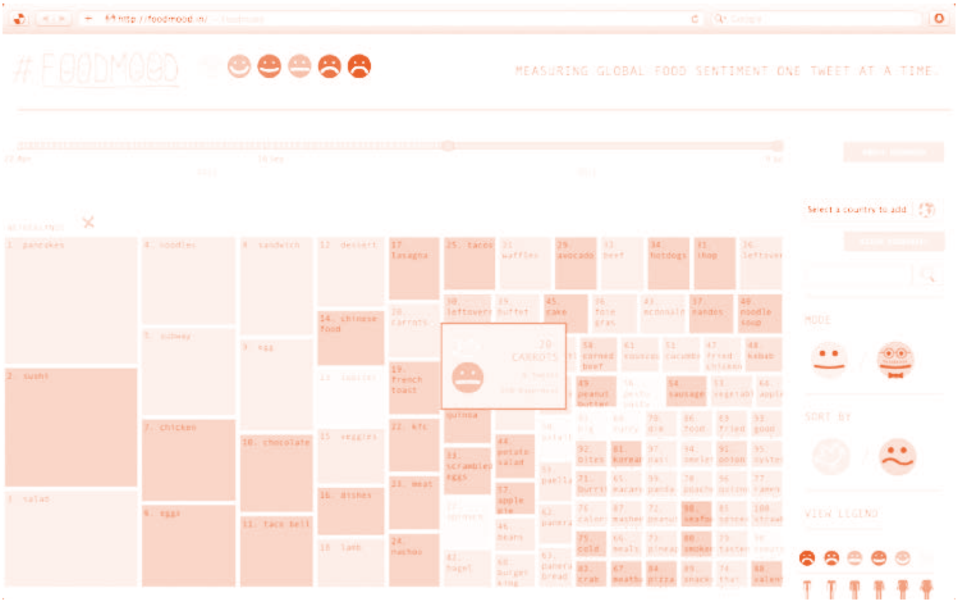
MentionMap is an example of a *network analysis tool* for Twitter conversations. MentionMap displays tweets from people, and charts mutual relationships. As an example, we entered the Twitter-ID of Ben Lorica, Chief Data Scientist at O'Reilly Media. This produced the following MentionMap:



The conversations in which Lorica has participated led us to The Economist, Microsoft Research (MSFTResearch), Stanford University, Berkeley student Neil Conway, Rafeboogs and David Wilson, among others.

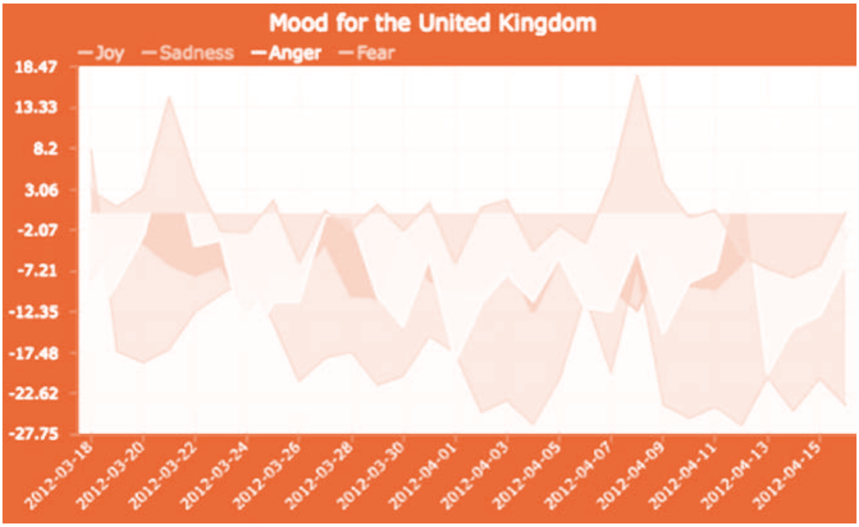
Sentiment analysis

FoodMood.in combines tweets with locations to measure the current mood with regard to food. The situation in the Netherlands is visualized below, with popular items such as pancakes, sushi, salad, as well as left-overs and carrots. Foodmood makes use of the sentiment classification of Stanford University, a 'trained' classifier that can cope with millions of tweets. Countries can be mutually compared, there is a list of the top-10 'happiest foods' of each country, and comparisons between the GDP of countries and the food scores are also possible.



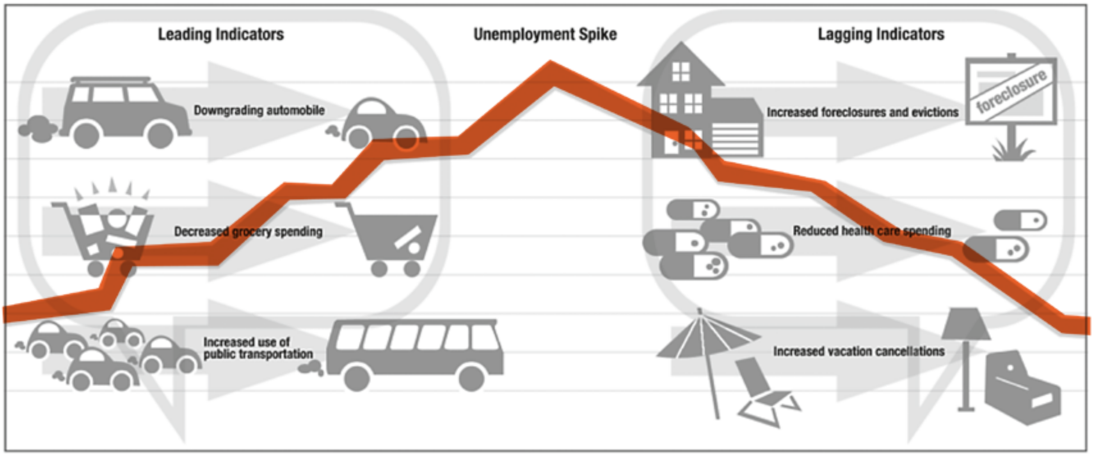
Besides general sentiment, we can also zoom in on individual tweets, so that the context becomes evident. An egg for breakfast, pancakes at a children's party, and chocolate on the sofa. We view the consumer in his or her domestic surroundings, as it were, and see how food is experienced.

In the UK, Bristol University measures emotions on Twitter and creates overviews of the situations:



Fear and sadness were predominant during the summer riots of 2011, and also later when the government announced that public spending was going to be cut drastically. The School of Informatics and Computing in Indiana built a model to predict stock market fluctuations on the basis of this kind of mood analysis. It investigated whether or not public sentiment correlated with the value of the Dow Jones Industrial Average. After a detailed text analysis, considering words such as calm, alert, sure, vital, kind and happy, the conclusion was that, in 86.7 per cent of the cases, the daily rises and falls in the final value of the DJIA could be predicted fairly accurately.

In October 2011, the American Federal Reserve Bank announced that it would follow consumer confidence in much the same way, via Facebook, Twitter, blogs, YouTube, forums, Associated Press, CNN and the Wall Street Journal. The Global Pulse project of the United Nations makes use of sentiment analysis among the population to predict unemployment and suchlike. In this context, SAS found six indicators in expressions on social media, including items such as converting to a smaller car or postponing a holiday.



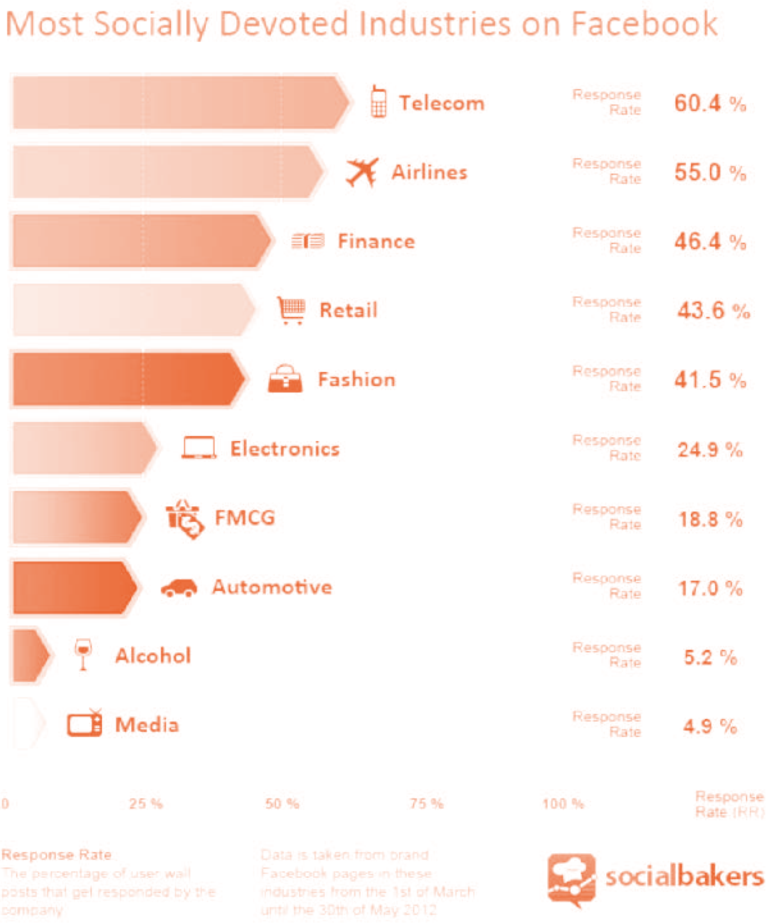
Analysis of social media using SAS shows increases in chatter about certain topics that are leading and lagging indicators of a spike in unemployment.

'SAS compared mood scores and conversation volume with official unemployment statistics to see if upticks in those topics were indicators of spikes in unemployment. The analysis revealed that increased chatter about cutting back on groceries, increasing use of public transportation and downgrading one's automobile could, indeed, predict an unemployment spike. After a spike, surges in social media conversations about such topics as canceled vacations, reduced healthcare spending, and foreclosures or evictions shed light on lagging eco-

> *nomic effects. Such information could be invaluable for policymakers trying to mitigate negative effects of increased unemployment.'*

11 The strength of Big Social Data

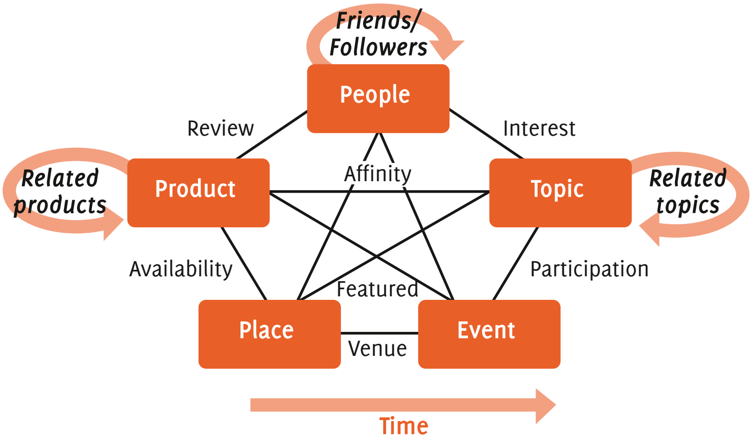
In this section, we conclude by aligning four concrete Big Social cases in a cross-section of various sectors: retail trade, banks, insurers, police, and product development in the computer industry. It may be too early to draw general conclusions about the sector in which Big Social could have most impact. If the activity of sectors on Facebook is an effective predictor of the impact, the range could look like this:



Independent of industry partitioning, we advocate a focus on processes and goals, as Altimeter presented in its above-mentioned report. The cases in this section concern cost-savings, product innovation, marketing and sales. Just as with the Target case in section 3, we see that a purposeful shift toward Total Data Management currently and immediately bears fruit – low-hanging fruit – when the approach is 'social'. The case of Walmart Labs shows that preferences displayed on social media online and offline can be directly converted to recommendations. The case of the British company Wonga illustrates that Big Data can make more bank products profitable. For insurers it is important to prevent fraud and to limit outlay for health costs. Finally, we examine the role of Big Social in combating crime.

1 *Walmart Labs & marketing innovation*

Walmart Labs, the social data R&D department of Walmart, originated with the acquisition of Kosmix, a social media analysis firm that is primarily known for its Twitter filter tool, TweetBeat. The most important achievement of the Lab has been the development of a tool that performs semantic analyses on Twitter, Facebook and Foursquare. In this way, the so-called *Social Genome* can be charted, consisting of rich profiles of customers, topics, products, locations, and events.



This concerns the interpretation of a network on the basis of relationships: a person is interested in a topic, a person attends an event, an event is related to a topic, an organization is associated with a product, etcetera. In this process, Walmart makes use of public data on the Web, its own data, and data from social media.

The first results are now evident. By means of data taken from Facebook and Twitter, better recommendations can be given on the website and in shops, using the new

Walmart app Shoppycat, for example, which provides gift suggestions to friends when someone's birthday is approaching.

2 *Borrowing money with Big Data*

Wonga, an English slang word for money, is a start-up that offers small, short-term loans without human intervention. On the basis of Big Data, Wonga cultivated this market and developed a profitable niche that had become unprofitable for the larger banks. People who need money often try to hide their sad financial state, but hard data does not lie, according to Wonga founder Errol Damelin.

It began with the experiment entitled *SameDayCash* and the beta-period ran as expected. For every successful loan, another defaulted. Nothing exceptional, but important facts could be gathered: about people and their behavior. SameDayCash fed the current Wonga algorithm. In the first year of operations, 100,000 loans were issued, with a collective value of 20 million British pounds. The algorithm examines all kinds of information sources, including social media. The system makes use of a basic set of 30 information points that are enriched with thousands of other data points. The algorithm exposes anomalies in the relationships. Because the loans are short term, there is a constant influx of new data. Financially this approach is feasible because it involves only small amounts.

3 *Fraud detection as a Big Social killer app for insurance*

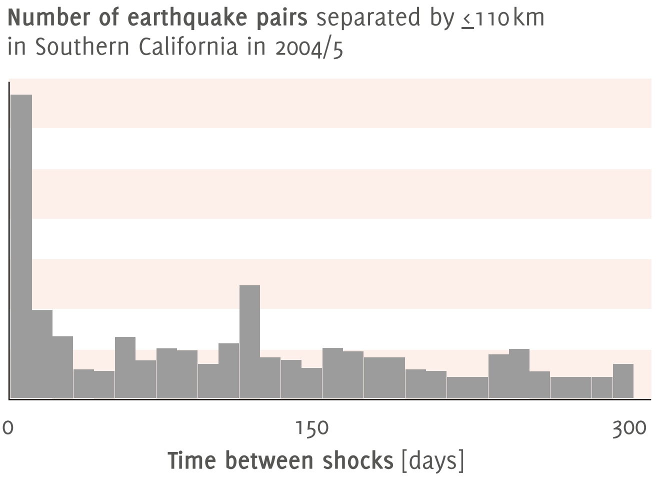
Fraud costs the American insurer Property & Casualty around 30 billion dollars a year: approximately 10 per cent of all claims for damages. By linking historical data to demographic profiles, the chance of fraud can be estimated. A person's network says much more. If three Facebook friends have already been caught, an extra check on a person may turn out to be quite smart. Thus, with richer profiles, Predictive and Social Analytics help reduce the most important expenditures of insurers, such as fraud and risk analysis. Text analysis also plays a major role in scrutinizing damage claim forms.

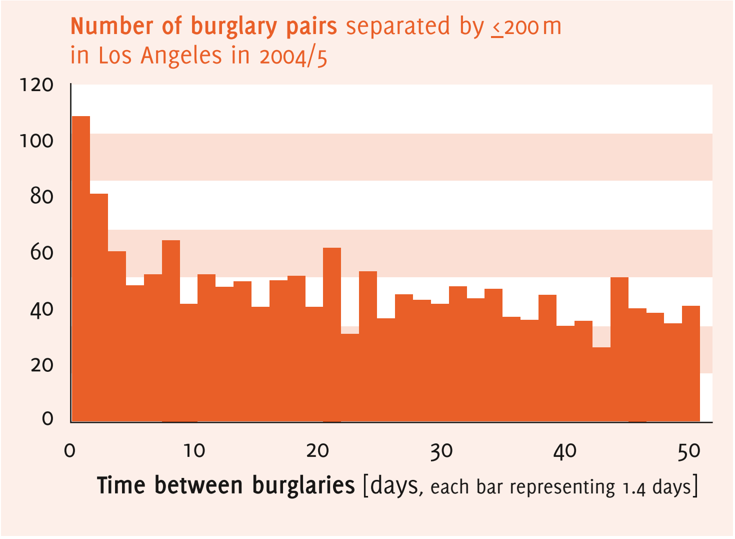
Traditionally, risk analyses may target social-demographic data, driving behavior, credit information and the like. New data points contribute to better risk analyses. Life insurance companies adjust their prices on the basis of medical data that suggest a healthy lifestyle. Many Facebook likes for extreme sports on a profile could perhaps lead to higher premiums in the future.

4 *Combating crime with the Predictive Police*

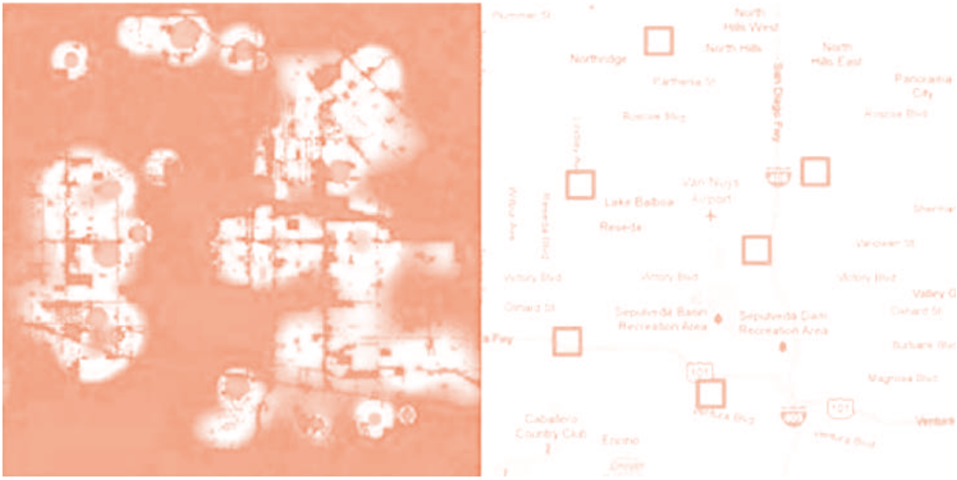
In the Netherlands, police officers go on duty with a smartphone in order to be able to pick up signals in the neighborhood from social media. In this way, they can show their faces before something serious happens in the schoolyard, for example. This kind of initiative is supported and shared by *Police 2.0*, a community that pursues the Big Social transition in their field.

In Santa Cruz, California, the police make use of an algorithm that helps prevent crime. The software was developed by two mathematicians, an anthropologist and a criminologist from PredPol, a 'predictive police' company, based on a model that predicts the aftershocks of an earthquake. Practice indicates that a criminal often returns to a spot where he or she has previously been active. *Aftercrimes* follow the same pattern as the *aftershocks* of an earthquake.





For some time now, various police departments have been making use of software by CompStat in order to predict crimes. These computer programs had only data at their disposal from at least a week old. The PredPol algorithms include real-time data instead. Taking into account location, time and type of crime, the software is capable of defining *prediction boxes* with an accuracy of 500 square meters.



In the first sixth months of the pilot project, the number of crimes fell by a quarter. It works quite simply: '*The suspect shows up in the area where he likes to go. They see black-and-white talking to citizens — and that's enough to disrupt the activity.*'

PredPol is an inspiring case but all kinds of traces on social media remain a valuable source too, for instance when riots occur like in some British cities in August 2011. Another good practice is to match police data with governmental sources. Even without algorithms a lot can be gained from smarter digital detective work: both in cyberspace and in the real world.

12 Summary and the organization of privacy

The adoption of plans in organizations for Big Data currently and predominantly covers the theme of Big Social: the customer side, inspired in particular by the social network activity of Web 2.0. But, if we take the concept of 'social' in a broader sense, an increasing amount of Big Data potential is released. This is more or less the route we have followed since the early nineties: first with Web Analytics, then with Social Analytics and now with Next-Generation Analytics. In this age of Big Data, further development is progressing toward Total Data Analytics and Total Data Management.

An important part of the discussion revolves around the extent to which organizations should embrace Big Social Data. The answer is: only on the basis of a well-grounded policy. Smart entrepreneurship in the growing dataflow is the key to capturing the raisins from the pie, so to speak. The question as to whether or not an organization initially is working with real Big Data (sets) is actually irrelevant. Scaling up will occur organically, and a good number of privacy issues are closely attached to this situation. We shall deal with these comprehensively inpart IV. Number four will be devoted to a Big Data roadmap, perceived and approached from various angles.

Modern Social Analytics applications enable organizations to understand the rhythms of human activity, to attach predictions to them, and to plan and implement corresponding actions: Understand, Predict & Act. The possibilities of personalization and hypertargeting are steadily increasing, and the toolbox is bursting at the seams. But do customers want that? It gives many of us a somewhat uncomfortable feeling to realize what commercial organizations know about individuals and groups. The organization of privacy and the guarantee of our personal integrity is perhaps therefore the domain *par excellence* to which attention should be paid. Big Data, Big Social and Big Brother are not worlds apart – certainly not in our human perception.

The beginning of this part was devoted to the issue of *What's Next in Big Data?* Many organizations are happy to be mere spectators at the moment, because there are no qualified best practices as yet and, this being the case, the financial risk could be considerable. But customers have a completely different mindset. Primarily they are alarmed by, for example, prospects of rising insurance premiums because they have presented themselves on the Internet a bit too enthusiastically, participating recklessly in certain leisure time activities, or showing themselves to be great fans of cigarettes and beer, to name just a few lifestyle choices. Regardless of what organizations may think about Big Data and Big Social, customers' Big Brother fear will force them to deal seriously with the situation, to adopt standpoints, and to express these vigorously.

Technology is advancing rapidly, we can make ever-better predictions, and we can step effortlessly from Web and Social Analytics on to Next-Generation Analytics. The accent is increasingly being placed on data and algorithms rather than on models. In short: the commercial power of Big Social Data is undeniable and is growing. At the very least, this entails increasing guarantees and responsibilities where themes such as privacy, personal integrity and, above all, perception and sentiment are involved. This is perhaps the very first observation for organizations to make with both feet firmly on the ground.

At present, Big Data is a dynamic and trending discussion theme. In the field of technology, of organization, of ROI, etcetera. For this reason, we are eager to keep in contact with organizations and individuals with regard to all 'next practices' that are currently being developed and evaluated: online at http://vint.sogeti.com/bigdata and, of course, in personal conversations.

Eight central Big Social definitions

As we have seen in this part, the realm of what we conveniently call Big Social contains different analytics and analysis flavors. For that reason this separate section is dedicated to the five main topics in this field. Eight definitions are provided in a logical order: one for Predictive Analytics, one for Sentiment Analysis, one for Web Analytics, four for Social Analytics, and one for Next-Generation Analytics.

The focus these days lies on Social Analytics. We start with the original and broadest angle by Lars-Henrik Schmidt, followed by the social media dominance of Kevin Roebuck, move on to the standard Gartner definition, and conclude with the converging social information and systems view by Mary Wallace. It is this converging way of looking at Big Social that underlies this this part.

The last definition concerns Next-Generation Analytics. It is open enough to accommodate all future Big Data development that is already explicitly appreciated in the definition of Predictive Analytics.

Predictive Analytics (Wikipedia)

'Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events. […] Technology and Big Data influences on Predictive Analytics: […] The volume, variety and velocity of Big Data have introduced challenges across the board for capture, storage, search, sharing, analysis, and visualization. Examples of big data sources include web logs, RFID and sensor data, social networks, Internet search indexing, call detail records, military surveillance, and complex data in astronomic, biogeochemical, genomics, and atmospheric sciences. […] It is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights. Today, exploring Big Data and using predictive analytics is within reach of more organizations than ever before.'

Sentiment analysis (Mejova, 2009)

‘As a response to the growing availability of informal, opinionated texts like blog posts and product review websites, a field of sentiment analysis has sprung up in the past decade to address the question What do people feel about a certain topic? Bringing together researchers in computer science, computational linguistics, data mining, psychology, and even sociology, sentiment analysis expands the traditional fact-based text analysis to enable opinion-oriented information systems. Sentiment analysis is closely related to (or can be considered a part of) computational linguistics, natural language processing, and text mining. Proceeding from the study of affective state (psychology) and judgment (appraisal theory), this field seeks to answer questions long studied in other areas of discourse using new tools provided by data mining and computational linguistics. Sentiment analysis has many names. It’s often referred to as subjectivity analysis, opinion mining, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion). [...] These are usually single words, phrases, or sentences. [...] Sentiment that appears in text comes in two flavors: explicit where the subjective sentence directly expresses an opinion (“It’s a beautiful day”), and implicit where the text implies an opinion (“The earphone broke in two days”). Most of the work done so far focuses on the first kind of sentiment, since it is the easier one to analyze.’

Web Analytics (Wikipedia)

‘The measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage.’

Social Analytics (Wikipedia)

‘Social Analytics is a philosophical perspective developed since the early 1980s by the Danish idea historian and philosopher Lars-Henrik Schmidt. The theoretical object of the perspective is socius, a kind of “commonness” that is neither a universal account nor a communality shared by every member of a body. [...] It might be said that the perspective attempts to articulate the contentions between philosophy and sociology. The practise of Social Analytics is to report on tendencies of the times.’

Social Analytics (Roebuck, 2011)

‘Social Analytics refers to the tracking of various media content such as blogs, wikis, micro-blogs, social networking sites, video/photo sharing websites, forums, message boards, and user-generated content in general as a way for marketers to determine the volume and sentiment around a brand or topic in social media.’

Social Analytics (Gartner, 2010)

‘Social Analytics describes the process of measuring, analyzing and interpreting the results of interactions and associations among people, topics and ideas. These interactions may occur on social software applications used in the workplace, in internally or externally facing communities or on the social web. Social Analytics is an umbrella term that includes a number of specialized analysis techniques such as social filtering, social network analysis, sentiment analysis and social media analytics. Social network analysis tools are useful for examining social structure and interdependencies as well as the work patterns of individuals, groups or organizations. Social network analysis involves collecting data from multiple sources, identifying relationships, and evaluating the impact, quality or effectiveness of a relationship.’

Social Analytics (Wallace, 2011)

‘If we look at the academic definition of Social Analytics “the process of measuring, analyzing and interpreting the results of interactions and associations among people, concepts, and facts” and apply this more broadly to the business, then a couple of things happen. Firstly we start to be able to harvest actionable social insights from existing enterprise applications, secondly we create a bridge that allows us to marry legacy business solutions with the new generation of social business platform, and thirdly we significantly increase the ROI we can realize from our social investment.’

Next-Generation Analytics (Gartner, 2010)

‘It is becoming possible to run simulations or models to predict the future outcome, rather than to simply provide backward-looking data about past interactions, and to do these predictions in real-time to support each individual business action.’

Part IV

Privacy, Technology and the Law

Big Data for Everyone through Good Design

Introduction

Reaping the fruits of Big Data

Predicting and targeting as the Big Trick

Data is the fuel of the digital economy. Everything that is possible nowadays may be extremely helpful and useful, but may also be threatening, or at any rate undesirable. The intelligence of a smartphone never goes amiss in the superstore when it comes to deciding what to buy for dinner, considering our dietary preferences. This is a wonderful thing of course, assuming that the collection and combination of data are dealt with transparently and discreetly. Quick digital advice supplied in this way, on the basis of preferences, is usually warmly welcomed. At one time Amazon started this in the retail business, to satisfy their customers and keep them satisfied as much as possible. The organization knows the customer and no one needs to feel cheated.

But in the case of the American Target chain – apropos of *targeting* – a clever analysis of purchasing behavior enabled predictions as to who was pregnant and also when the delivery was likely to take place. But Target was not transparent to the customer with respect to this kind of practice, and so it happened that the father of a teenager was unpleasantly surprised by the offers Target made to his daughter. True, Target was right and the girl had indeed been pregnant for as many weeks as they said, but it initiated an extensive discussion in the press about what organizations know about us and how they are using their Big Data knowledge to sell as much as possible.

Appropriate dead-on targeting is a wonderful thing, but we are rarely told what organizations know about us and how that knowledge is being used. For decades, this transparency has been a crucial part of the so-called *Fair Information Practice Principles* (FIPS, *Data Protection Principles* in Europe), but obviously some corners are being cut here. You may be familiar with the Target example from part III *Big Social* and there are those who will make no issue of it; but what if your credit, mortgage or insurance application is rejected because the data indicate that your financial situation and/or health constitute an unacceptable risk to the provider? Which information is involved here? Where does it come from? How has it been collected? Do you have access to it? Can you change it? These simple and fundamental questions are a tricky problem in the age of digital information and have been for decades, in fact. Countless books have been written and there is much jurisdiction on the positively Kafka-esque examples of people who have come to be put in a bad light.

By and large, the major advantage of Big Data is its ability to make better predictions and selections. To organizations, the opportunities are ubiquitous: fraud detection, more efficient energy supply, offers tailored to the customer, anticipating epidemics, etc. One would expect that this would benefit all, but which data are being gathered by digital monitoring systems and what is being done with them? Do we have any idea about them and any control over them? The consumer/citizen, who is kept in the dark, often experiences this as one Big Trick. He or she feels that his/her privacy is being invaded – and so it is, of course. This part, *Privacy, technology and the law*, addresses that confrontation.

Transparency, choice and Privacy by Design

Where does this lead us? First and foremost to the conclusion that any organization engaged in Big Data should be thoroughly familiar with privacy and data protection. Evidently this is not always the case. If we are transparent and open about what we are doing, if customers are offered a clear choice whether or not to supply information and if we are implementing the *Privacy by Design* principle, the three most important steps have thus been taken. It is along these lines that this part seeks to contribute to a fundamental privacy awareness that structures business activities in terms of transparent data management, which is for the benefit of customer and supplier alike.

There can be no two ways about it: whoever tries to find an on/off switch for privacy will never find it. It is much more important to choose the right direction, so as to exploit Big Data to its full advantage. What we should aim for is “Big Data gain for everyone.” The best way to go about this is by recognizing the privacy issues, exposing them in detail and leading them in accepted directions in all candor.

Big Data gain for everyone

There is no more forceful way of putting it than Meglena Kuneva did: “Personal information is the new oil of the Internet and the new currency of the digital world.” Ms Kuneva, EU Commissioner for consumer protection, said this during her keynote speech at a roundtable meeting on data collecting, targeting and profiling in Brussels in late March 2009. But broadly speaking, it is about digital data, online as well as offline.

Ms Kuneva outlined the situation as follows: “The boom in terms of volume of all the collected personal data and its use for commercial purposes is one of the most important and controversial issues in the rapidly changing world of digital communication.”

Boom, volume, speed, economic value and various kinds of digital personal information: welcome to the Big Data era. When using these terms, we find that digital privacy is completely analogous to the concept of Big Data, which is defined as a combination of *Volume*, *Variety* and *Velocity*, supplemented by some with *Veracity* and *Value*. This is important and controversial: a godsend to customer service, but with all the usual challenges.

"The Internet," said Meglena Kuneva in 2009, "and the new generation of digital communication and digital platforms offer huge opportunities to consumers. In terms of choice, access and opportunity they are among the most empowering tools consumers ever had access to. [...] Obviously we want these new opportunities to evolve on a permanent basis and therefore we need to boost people's confidence, which will be conducive to their participation." Kuneva emphasized that "the Internet is largely an advertisement-driven service and is kept going by the development of marketing on the basis of profiles and personal data."

She added the following comment: "Over 80% of the young Internet users think that all kinds of personal data are being used and shared in one way or another without their permission, and actually this is true." When it comes to privacy protection, Ms Kuneva feels that the sorely-needed solution is to be more transparent when it comes to collecting data. "Consumers need to be informed that their data are being bought and sold, and they ought to be offered the opportunity to supervise these activities themselves."

These views are by no means new. We have known for some decades now that the privacy landscape is very much in the making, thanks to increasing digitalization. Concisely paraphrased, the introduction to the book *Technology and Privacy: The New Landscape* (1997) puts it as follows:

> *Digital privacy is the capacity to negotiate socio-economic relationships by controlling your own personal information. Rules and regulations, policies and technology increasingly structure people's relationships with organizations and governments.*

There are huge differences in terms of privacy regulations both nationally and supranationally, but we all tend to agree on one thing: the enormous potential of the digital economy. It is extremely important to harmonize data protection and business interests, and lay down their interrelationships in the various legal systems in a coherent manner.

Big Data to become more privacy-friendly

Big Data is not privacy-friendly. In November 2012 Brendon Lynch, Chief Privacy Officer of Microsoft, emphasized this once again at the European Data Protection Congress of the International Association of Privacy Professionals (IAPP). Even when anonymization has taken place, certain core data have been deleted or data have been "scrambled," it is still perfectly possible to link specific information unequivocally to an individual, to a computer or some other personal device on the basis of the links in different Big Data collections, online as well as offline.

To counteract this *linkability* and (re)identification, Microsoft has now operationalized a technological Privacy by Design solution – after years of development – that guarantees the quality of digital data for targeting by organizations, while making individual people untraceable with absolute certainty. In Big Data circles the method is known as *Differential Privacy*.

Target, for one, had never bothered about notice and consent, two important privacy principles, but Brendon Lynch asked himself: how can you expect everything that happens in a Big Data world to be reported in detail, and that explicit consent be asked? In the same way that the financial world has its flash transactions, our Big Data world is absolutely full of flash information.

Interestingly, a visionary historical quote on the American Privacy Rights Clearinghouse website makes unequivocally clear that Big Data recombination is at the heart of privacy concerns today. Back in 1977, ominously two hundred years after the first official copy of the Declaration of Independence was printed, the US Privacy Protection Study Commission already exposed the "real danger" to come in the following terms: "the erosion of individual liberties through the automation, integration, and interconnection of many small, separate record-keeping systems, each of which alone may seem innocuous, even benevolent, and wholly justifiable." The crucial difference is that in our Big Data age the number of record-keeping systems has exploded and they are huge instead of "small," as goes for the continuous monitoring and real-time analysis and recombination of XXL data streams containing Personally Identifiable Information. As stated in the Obama Government's Consumer Privacy Bill of Rights of 2012, consumer data privacy is the lubricant and fuel for the global digital economy. So let's keep the engine running! And, by the way, when personal data is fueling the economy, why not tax it, as was suggested in for instance France?

The personal information economy

The so-called "personal information ecosystem" is described in the *Protecting Consumer Privacy in an Era of Rapid Change* report by the American Federal Trade

Commission of March 2012. This document contains extensive recommendations to organizations and policy makers along the lines, successively, of Privacy by Design, simple choices for consumers and transparency. In the economic system, the individual is central and a wide variety of data is being collected about all of us. This is being done by media, government institutions, energy suppliers, airline companies, credit organizations, the retail sector, telecom companies, cable companies, insurers, banks, hospitals, doctors, drugstores, browsers, commercial websites and social networks. Information brokers, including credit companies and the advertising industry, use and combine these data and in this way they end up in banks, for example, or with marketers, media, authorities, legal organizations, individuals, upholders of justice and employers. It concerns online and offline data, originating from individuals, their computers or other devices. The only situation where the recommendations of the FTC report are not applicable is where an organization only collects privacy-neutral information of under 5,000 individuals a year, and does not share it with third parties in any way whatsoever.

It may remain a mathematical limit, but with the increasing demand for Privacy by Design, transparency, openness, notice, consent and more particularly the individual control of the information collected with regard to storage, processing, combining and dissemination will increasingly assume concrete shape. Organizations need to be aware of this and be ready. This means: gathering fundamental knowledge and organizing your operation accordingly.

Privacy and trust are the lifeblood of digital business

To succesfully drive a car in this digital age we can rely on advanced navigational services but there really is more to that than just mapping the route. At the Sogeti Executive Summit 2013, Simon Hania, TomTom's Chief Privacy Officer, made this very clear. His talk was on location services and privacy or using geolocation in a trustworthy and compliant way. Mr. Hania discerns four overlapping digital trends that tend to threaten privacy and trust: cloud computing, location services, the Internet of Things and Big Data. You all find these nowadays in what we call the Connected Car and the Cloud-Connected Car.

TomTom is in the business of revolutionizing navigation with layers of information. Of course there are the base maps and for a start people can share them. TomTom focuses on allowing drivers to take the most efficient route based on proprietary IQ Routes and HD Traffic technology. Today, TomTom Traffic covers 99.9% of all roads. To create their services TomTom captures data from various sources: in-dashboard GPS, fleet GPS, app GPS, detector loops and cameras, and GSM among others. This is where the issue of privacy and trust comes in.

TomTom operates a huge trip archive with anonymous location and speed information from their community. Each day, five billion speed measurements are being fed into the system, and it now contains five trillion measurements donated by customers that drove 50 billion kilometres, visiting every spot over a thousand times. For instance the exact travel time to a hospital may serve as a reality check that can help save lives, since there is a significant difference between a route based on theoretical maximum speed and real-world speed measurement. TomTom tracks where customers are coming from, what routes they take, the amount of drivers passing, and combines these data with other geo-based information sources for additional analysis.

All over sudden in April 2011 rumor had it that TomTom would share Personally Identifiable Information (PII or personal data) with the police without their customers knowing. The company was eventually cleared of all data violation allegations but since this unfortunate incident communication around privacy and data protection has become a key priority. Informing users must be fully explicit, including opt-in. TomTom only uses community input with permission and is in the business of profiling roads and routes, not people.

Companies should take the following seven Ps into account and be totally transparent in these contexts: Principles, People, Policies, Projects, Processes, Procedures, Paperwork. The vision: community input or crowdsourcing is strategic, and privacy helps to realize business objectives by ensuring trust, it being an integral part of business continuity above and beyond legal compliance.

These six simple privacy questions are leading:

1. What data are we processing?
2. Why are we processing personal data?
3. When can we destroy the personal data?
4. Who will have access and will be accountable?
5. Where will we process and store the personal data?
6. Will we have a legitimate basis for processing?

Avoiding re-identification based on Personally Identifiable Information (PII) is key. For TomTom this means that it uses its historic trip archive only for road, traffic and related purposes; that there will be no access to raw data outside of TomTom; and that there is sufficient aggregation to make re-identification impossible. A ‘privacy czar’ should be appointed in every company to oversee that so-called Privacy by Design is being developed, implemented and controlled in the right way.

1 An anatomy of Big Data anxiety

1.1 Digital intangibles are terrifying

In the past few decades, what had been owned and physically possessed by people for thousands of years – the mine and thine, private behaviors and domains over which no one else had any say except when expressly invited – has now shifted to digital information: to all sorts of personal data in databases and our everyday activities and dealings on computers and online. In short, the concept of possession also covers our digital *Personally Identifiable Information* (PII), and the ownership, access, collection, storage, use and dissemination of this information.

The more digital data going around in all shapes and sizes, the greater the anxiety that, for one reason or another, this might result in a situation where far more private information becomes known than we would like. This may vary from video pictures, location-oriented information and social media to databases, surfing and purchasing behavior on the Internet, as well as the data that smart energy meters can collect nowadays. The possible linking of these and other data – in other words, actual Big Data use in all its aspects – is not yet transparent enough, and there is justifiable anxiety concerning the effective protection of information.

This is demonstrated by hackers and cyber criminals who manage, time and again, to penetrate into all sorts of digital systems – subsequently plundering bank accounts, reselling information, or simply putting it online so that everyone can access it. The way in which the digital arms race between attackers and defenders is going to develop is largely occurring beyond our range. This, too, is cause for concern and, in the absence of facts and a sound risk assessment and considering the security leaks that keep on occurring, it continues to nourish anxiety and speculation.

With regard to privacy, "digital" has largely taken the place of "physical." We may put on as many dark glasses as we want, but our digital traces tell a great deal about us, and these traces are relatively easy to get hold of if you want to. That is the situation as we know it today, or at any rate, as it is perceived. And the various actual situations as well as their perception are ripe for clarification and improvement. An example…

In late November 2012, the Dutch TV program *De Wereld Draait Door* ("The World Is Turning Mad") called attention to the Electronic Health Record (EHR). After earlier opposition, this facility will be started up again in 2013 in the form of an opt-in arrangement called the Personal Health File, which means that people have to give their explicit permission to be included in the register.

Wilna Wind, director of the Dutch Patients' Consumer Federation, and Internet expert Alexander Klöpping were sitting opposite one another. Wind is a passionate advocate of the EHR, whereas Klöpping is vehemently opposed to it. He alarmed the audience with stories of his experiences with the hacker scene. At the end of the discussion, Matthijs van Nieuwkerk, the TV host, asked the audience who would still consider joining the EHR. No one raised his or her hand. Despite the new opt-in regulation, people are still quite apprehensive – not least because Ms Wind repeatedly assured the audience that EHR security would improve from a score of 4 ("unsatisfactory") to a score of 8 ("good") within six weeks, while everyone was fully aware that this discussion has been going on for years.

What can one conclude from this? Is Klöpping right? Evidently we hate to take the risk. To start with, weak spots in our privacy and data protection must be repaired as well as possible with the help of combined technology, procedures and regulation. We should aim for a so-called structural "Privacy by Design": privacy and data protection designed in conjunction with services and practices: an approach that is easily explained, offers optimum security, and inspires confidence.

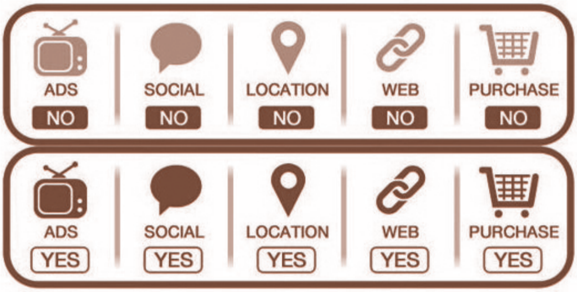
Currently, specific fields of application for the Privacy by Design approach are the following so-called potential "Privacy-Invasive Technologies" (PITS). Obviously health care and Big Data Analytics are also included in the list:

1. Camera surveillance
2. Biometric recognition
3. Smart Meters and the Smart Grid
4. Mobile devices and communication
5. Near Field Communications (NFC)
6. RFID and sensors
7. Redesigning IP Geolocation Data
8. Remote Home Health Care
9. Big Data and Data Analytics

http://www.privacybydesign.ca

It is this trio of explanation, security and confidence, along with responsible behavior on the part of organizations and individuals, that will have to help us cope with our well-grounded – as well as irrational – fear of loss of privacy. When everyone understands the ins and outs of the matter and how the developments are likely to work out, this may form the basis for renewed consideration of a trade-off of personal data with an eye toward better individual service.

Fact and perception with regard to privacy can be communicated and addressed excellently through accreditations, quality labels and easy-to-read attachments. In 2012 the American Association for Competitive Technology, among others, made the set of pictograms shown below, which indicate what happens and does not happen with our personal data in the mobile apps that we download on our smartphones and tablets.



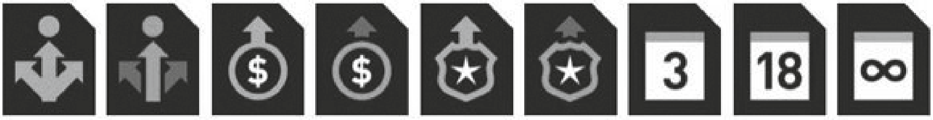
In America, the AppRights movement is working on a private member's bill, *The Application Privacy, Protection and Security (APPS) Act of 2013*, which is to regulate the collection of data via mobile devices and apps. In September 2012, the U.S. Federal Trade Commission (FTC) already issued a clear set of guidelines for app developers called "Marketing Your Mobile App: Get It Right from the Start." Only to help them "comply with truth-in-advertising standards and basic privacy principles."

Fact is that many mobile app makers leave consumers confused or in the dark when it comes to app privacy options. Even worse, they deliberately mislead people, thus drowning the Golden Opportunity of monetizing Personally Identifiable Information in FUD: fear, uncertainty and doubt. Therefore, the FTC explicitely warns: "Laws that apply to established businesses apply to you, too, and violations can be costly."

To keep themselves out of trouble, app owners and marketeers should adhere to well-known Fair Information Practices regarding "Truthful Advertising" and "Privacy."

As from March 14, the European Union is moving in the same direction. The European data protection authorities, gathered together in the so-called "Article 29 Working Party," recently have detailed the specific obligations of app developers and all other parties involved in the development and distribution of apps under European data protection law. Other parties include app stores, advertising providers, Operating System and device manufacturers. Special attention again is being paid to apps targeting children.

This is happening more and more. Mozilla, among others, uses pictograms that indicate, for example, whether a website shares or sells data, passes them on to a government agency without a court order, and how long they are stored.



1.2 Internet and privacy are uneasy bedfellows

A survey among the American population of 1997, when about one quarter of all Americans were online, showed that, even then, people were extremely worried about Internet privacy. In the same year, the *Framework for Global Electronic Commerce* of the Clinton administration put it as follows:

> *Americans [and all other people] treasure privacy, linking it to our concept of personal freedom and well-being.* ***Unfortunately, the GII's [Global Information Infrastructure] great promise – that it facilitates the collection, re-use, and instantaneous transmission of information – can, if not managed carefully, diminish personal privacy.*** *It is essential, therefore, to assure personal privacy in the networked environment if people are to feel comfortable doing business.*

The Internet should not be a playground for undesirable and improper behavior, as this is detrimental to economic potential, or so it was argued. This results in a lack of confidence, causing customers and providers to stay away and preventing the free world market from developing as it should.

We are eager to allow the social and economic potential of the Internet to flourish as an everyday part of our lives. But its openness and speed carries a huge inherent risk of misuse. All stakeholders involved must accept responsibility here, ideally with the private sector taking the lead, as it has the greatest economic interest.

In 1997, people expected more and more individuals and organizations to actively use the Internet if their privacy could be fully guaranteed. But despite all such privacy concerns, the Internet continued to boom – so saying and doing are apparently two different things. Therefore, may the Dutch be expected to join the new EHR in due course, in spite of all their present skepticism?

To what extent is the fear of an EHR and other Big Data initiatives realistic? How easy is it to turn a score of 4 for security into an 8? Perhaps that is not hard at all and after all, people's reaction to change is often based on gut feelings.

The issue of whether or not current emotion and misgivings with regard to personal data will eventually wane again requires further analysis. The EU, at any rate, feels that generally speaking online privacy is not properly regulated:

> *Internet privacy is not properly protected. This is the view of the European Commission, which has drawn up new rules. However, these may not come into effect until 2015. In advance of this, Dutch regulations will be tightened considerably as of this year. As they should, because the present law dates from 1995 and is hopelessly out of date. The major changes are:*
>
> - *Data of consumers must not be used without their explicit permission.*
> - *Companies have to outline their privacy policy in plain terms.*
> - *Consumers are given the so-called "right to be forgotten."*
>
> *Companies not complying with the rules risk fines of up to 2% of the turnover, which for large businesses may amount to tens or even hundreds of millions of Euros.*
>
> RTL Z, 14 January 2013

It is the intention to ratify the so-called "General Data Protection Regulation" in the European Parliament in 2015. We are no longer talking about a guideline for national legislation, but a new European "act." This means that data-processing organizations will have to meet stringent requirements. Fines for businesses may run to 2% of the turnover.

1.3 Reasonable anxiety

Anyone compiling an anatomy of fear of Big Data will find that this anxiety is well grounded. In the previous century, the first Big Data factories, the credit-rating companies, followed rather dubious practices. They disregarded the law, made few or no rectifications of mistakes in data, combined all sorts of databases in a creative fashion in order to gather as much personal information as possible, and were constantly involved in court cases and hearings due to their procedures. In 2004 Robert Ellis Smith published a retrospective on the subject entitled *Ben Franklin's Web Site: Privacy and Curiosity from Plymouth Rock to the Internet.*

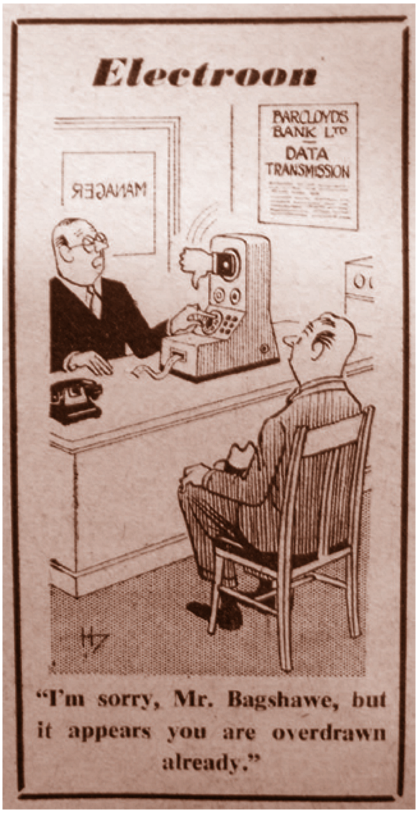
American credit-rating agencies such as Equifax, Experian and TransUnion combined citizens' personal data from a variety of databases, linked them with social security numbers, and subsequently used and sold those profiles. These credit-rating companies furnished information to banks and authorities that had to make decisions with regard to car loans, life insurances or benefits. According to Robert Ellis Smith, they also resold the information.

A negative credit assessment might ruin you. This is what happened to Keith and Phyllis Mirocha, who were not given a mortgage for their new home although they were clearly the victims of mistaken identity. With all the goings on and the legal battle they had to enter into – with TransUnion, in this case – the couple lost their jobs into the bargain.

The Mirocha case has many elements that nourish the Big Data anxiety even today:

- an unequal fight: large institutions versus the common man
- information is used without permission
- systems are making decisions without human intermediary.

It is quite a job to be proved right in this type of case. Even after the mistake in the file of the Mirochas had been detected and Trans Union had promised to correct the data, they were still refused a mortgage. The problem was that the wrong information was still in the system, and this barred the loan. It was a disagreeable situation that looked like something that had emerged from the hilarious "Computer Says No" sketch from the BBC series *Little Britain*. The following cartoon from *Electronics Weekly* of 2 November 1960 shows that this situation has a long history. Note the thumbs-down signal so emblematic of Facebook nowadays. It almost looks like a reference to the plan of Schufa, the German credit-rating agency, to link personal information from Twitter, Facebook and LinkedIn to their 66-million-customer database for the sake of better credit profiles.



"I'm sorry, Mr. Bagshawe, but it appears you are overdrawn already."

Two specific Big Data-related developments can be added to the three anxieties from the days of the Mirochas:

- Digital data reach other parts of the world in no time at all. Possible privacy invasion by America, in particular, are anathema to the Europeans.
- Personal information shared by people on social media threatens to be used against them by governmental authorities, insurers and other organizations.

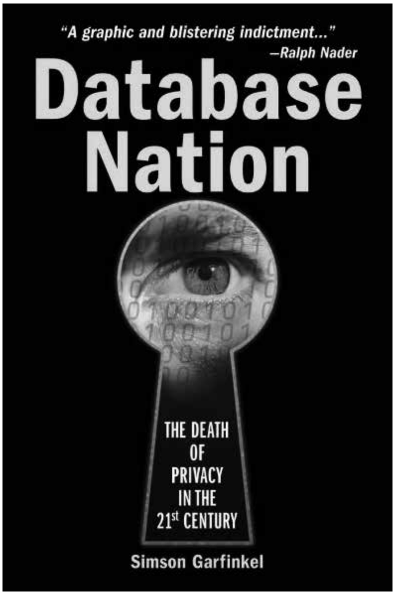
The "Computer Says No" syndrome has remained a controversial issue to date. This is clearly illustrated by the first *Issues Paper* published by the South Australian Law Reform Institute of Adelaide University Law School in May 2012. Its title is: "Computer says no: Modernisation of South Australian evidence law to deal with new technologies."

The story of the Mirochas is illustrative of what was going on at Equifax on a large scale. During a hearing it turned out that employees had been pressured to obtain a certain quota of negative reports on consumers. This put them up to fabricating data in creative ways.

Equifax was told by court order to bring the guidelines for the proper use of information to the notice of their staff, but that judgment was disregarded for years. To a considerable extent, this supports the fear that the law is powerless. Flagrant violations of confidence and sentiments such as “they will do what they like, no matter what, and who is to stop them?” are running rampant.

1.4 Fear, uncertainty & doubt

The fear of privacy loss as a consequence of large-scale application of technology was further fuelled by books and investigations covering violations of privacy and the ways and means to do so. Particularly *The Naked Society* by Vance Packard (1964) was responsible for the sentiment, followed by two influential publications about Big Data *avant la lettre* by Alan Westin: *Privacy and Freedom* (1967) and *Databanks in a Free Society* (1972).


“A graphic and blistering indictment...”
—Ralph Nader
Database Nation
THE DEATH OF PRIVACY IN THE 21st CENTURY
Simson Garfinkel

Eventually, after *Database Nation: The Death of Privacy in the 21st Century* (2001) by Simson Garfinkel, the general public was completely confused. FUD, the familiar *Fear, Uncertainty & Doubt*, had definitively become the standard. At the same time, this was the main point of the criticism. Was it true that nothing at all was being done about this kind of practices? We can quite well understand that people tended to respond to fears and facts from the past just to be on the safe side, but what is the actual truth?

Four lessons learned from the widespread fear during the early Big Data era

1. Without pressure from the outside, organizations are unlikely to change their behavior. Corrections can be effected through public fear and agitation.
2. The real fear concerns the improper use of data by third parties, especially if this happens deliberately, as in the case of data trade.
3. Eventually, indignation about the practices of American credit-rating companies resulted in the 1973 *Code of Fair Information Practices*, followed by the *Swedish Data Act*. Both are based on five principles of proper management of personal information, i.e.:
   - There must be no secret data collection.
   - People should be able to check what has been collected about them and how that is being used.
   - Use of data for other purposes is only allowed after the person in question has given his/her permission.
   - A person should be able to correct or amend his/her Personally Identifiable Information (PII).
   - Every organization creating, managing, using or disseminating PII, has to guarantee the reliability of those data for the intended purpose and make sure the information cannot be misused.
4. Another lesson to be learned from early Big Data history is that valuable time passes between the introduction of rules and regulations and the actions based on such rules and regulations. Therefore it is reasonable for people to be anxious about some parties having an opportunity to ignore the law for a longer time before mending their ways.

1.5 Privacy by Design as solution track

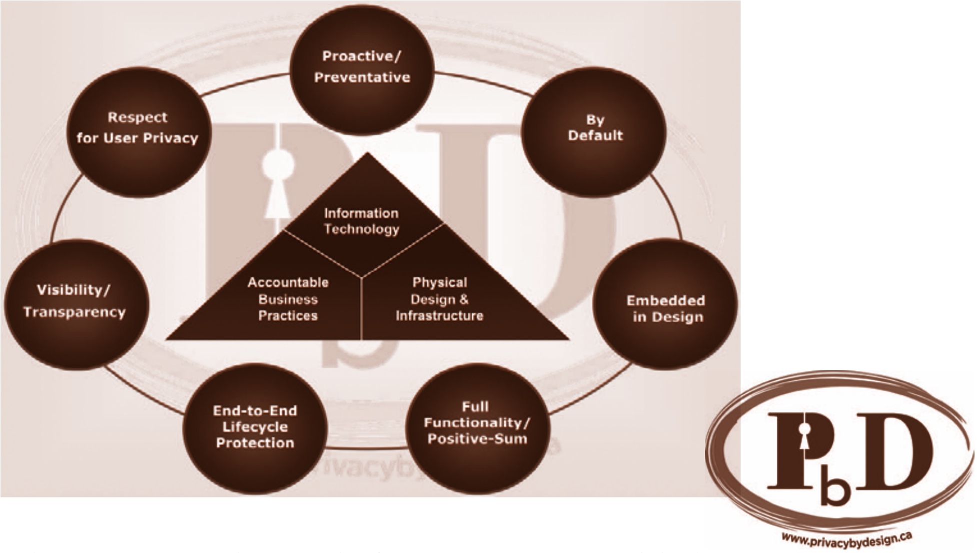
Anyone would think that institutes are focusing less on personal security and privacy than on the business opportunities and efficiency gains offered by new technology. Personal security is not naturally embedded in the system at the outset – it follows at a later stage. It took quite a while, for example, for credit card companies to start texting a verification message after a transfer.

In *Unsafe at Any Speed* (1965) the activist Ralph Nader, who also wrote the foreword to *Database Nation* (2001), analyzed the lack of interest with regard to personal security on the part of the automobile industry. The subtitle of the book is *The Designed-in Dangers of the American Automobile*, but dangers and negative effects need to be neutralized all along the line by building in countermeasures, Nader says:

> *A great problem of contemporary life is how to control the power of economic interests which ignore the harmful effects of their applied science and technology.*

Safety for the driver, passengers and the environment – it is all part now of the design and therefore of the business. One might say that Privacy by Design, also called the Golden Standard, is the safety belt, the cage, the airbag and the particulate filter of the Big Data business. Nowadays these things are built in at the outset. Privacy by Design is the ideal way forward when it comes to the simultaneous designing and adaptation of technology, procedures and regulations, with a view to assuring optimal safety and guarantees. The harmful effects and risks of driving a car have not been entirely removed, but certainly diminished to a great degree.

Likewise, the interest of stakeholders in the "safety" of personal data will continue to grow. Security and control should be an integral part of the design of systems as well as the eco-system within which they are functioning. This is conducive to the win-situation for all parties and the flourishing of economic models and opportunities, as was earlier explained by the Clinton administration in its *Framework for Global Electronic Commerce.*



As privacy, data protection and personal information represent such high economic and relational value, Ann Cavoukian, the Canadian Information and Privacy Commissioner and "mother" of Privacy by Design, proposed these seven basic principles around the core of each organization, i.e., technology, design and infrastructure, and the operation itself:

1. Privacy by Design means that you take proactive and preventive action: not reactive – no repairs afterwards.
2. Privacy guarantee needs to be the default setting.
3. Privacy needs to be embedded in the design.
4. Go for full functionality: not a poor trade-off but a clearly positive balance.
5. Solutions need to be totally conclusive and unequivocal: end-to-end security at all times.
6. Ensure full visibility and transparency: openness is your leitmotiv.
7. Deal with privacy respectfully: particularly by focusing attention on the individual.

These principles are further operationalized in the conclusion of this part and you are reminded of them by the Privacy by Design (PbD) questions in the margin.

1.6 Our landscape of technology and privacy in a nutshell

The book *Technology and Privacy: The New Landscape*, which was published over fifteen years ago, contains an apt definition of digital privacy. The ensuing fear and hope are also touched upon, and the concept of Privacy by Design is also put forward from converging perspectives, albeit before the term truly existed:

> *Privacy is the capacity to negotiate social relationships by controlling access to personal information. As laws, policies, and technological design increasingly structure people's relationships with social institutions, individual privacy faces new threats and new opportunities. [...]*
>
> *The essays in this book provide a new conceptual framework for the analysis and debate of privacy policy and for the design and development of information systems. The authors are international experts in the technical, economic, and political aspects of privacy; the book's strength is its synthesis of the three.*

Here we give a brief explanation of a number of central concepts (we also refer to "Literature and Illustrations" at the end of this book).

Privacy-Enhancing Technologies

Technology and Privacy: The New Landscape *contains a chapter by Herbert Burkert entitled "Privacy-Enhancing Technologies (PETS): Typology, Vision, Critique." This emeritus professor is currently in charge of the research center for Information Law at the University of Sankt Gallen, Switzerland.*

Privacy-Invasive Technologies

One year later, in 1998, the Australian e-business consultant Roger Clarke placed the abbreviation PITS – Privacy-Invasive Technologies – opposite PETS. An up-to-date overview can be found on the PET wiki of the Center for Internet and Society.

Dataveillance

The Dataveillance & Information Privacy *pages of Roger Clarke provide an interesting overview of PITS, PETS and their context. The term* dataveillance *was coined by Clarke. He discussed the concept in the article “Information Technology and Dataveillance” in the* Communications of the ACM *magazine of May 1988. In addition to* surveillance *and* dataveillance *you may nowadays also come across the terms* sousveillance *and* uberveillance.

PETS and Privacy by Design

Recent literature on PETS and Privacy by Design:

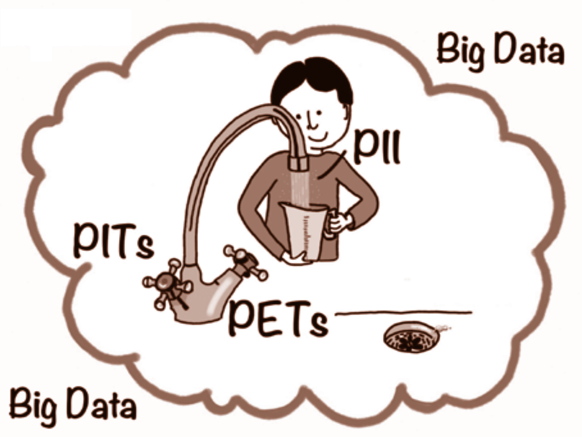
- *the* Handbook of Privacy and Privacy-Enhancing Technologies *(2003), devoted to intelligent software agents;*
- Privacy-Enhancing Technologies: A Review *by HP Laboratories (2011);*
- Privacy by Design in the Age of Big Data *by the Canadian Information and Privacy Commissioner Ann Cavoukian and IBM’s Big Data guru Jeff Jonas (June 2012);*
- *the report* Operationalizing Privacy by Design: A Guide to Implementing Strong Privacy Practices *by Ann Cavoukian (December 2012).*

Privacy by Design and PETS are in a process of rapid development

The relation between Personally Identifiable Information (PII), PITS, PETS and Privacy by Design is very much in the making. A critical view is provided by the article “Regulating Privacy By Design” (2011) written by Ira Rubinstein, a Senior Fellow in the Information Law Institute of the Center for Democracy and Technology, among other things. Rubinstein has doubts about the worldwide enthusiasm with which Privacy by Design and PETS have been greeted in recent years. The thing is that new worlds are hidden behind these concepts and this is where the work really begins, amidst rapidly developing technologies and data flows.

During the last few decades, a lack of clarity about and mistakes in the collection, storage, use and dissemination of personal information have given digital technologies – which enabled all this in the first place – a reputation of Privacy-Invasive

Technologies (PITS). To prevent us from having to undergo a ruthless cold shower, we installed a faucet of rules, so to speak. In this way the privacy invasions of cold PITS could be dosed and, by adding hot water in the form of PETS, the water is now no longer ice cold but pleasantly warm. The regulated dataflow that we collect in a measuring jug stands for our Personally Identifiable Information (PII). If too much of it is tapped off, or if it turns out to be a cold shower after all, then down the drain it will have to go – which leaves us with the choice of whether or not to try again. The PII water serves to irrigate the economic relationship with all kinds of service providers.



René Speelman, 2013

It is a striking analogy and indeed: PII, PITS, PETS and the companion Privacy by Design collectively form the basis to doctor privacy security, the aim being to prevent privacy invasions with the help of advanced digital technology.

PETS and Privacy by Design are a major supplement to the original "well" of Fair Information Practice Principles (FIPS), which have no explicit affinity with technology. The development of the technology-oriented Privacy by Design, and consequently, the PETS combined with transparency about and options within business practices and information systems, is a necessary Total (Personal) Data Management approach. All conceivable stakeholders in and outside organizations need to be actively involved and take their *Don't Be Evil* responsibility.

This is why Ann Cavoukian, among others – the "mother" of Privacy by Design – keeps going on about openness and transparency. The goal is "PII for everyone" in a sound economic context. Of course, smart technological PET solutions such as Differential Privacy should be an integral part of this. However, as in the case of PITS, the unbridled dataflows of smart energy meters, for example, and of consumption meters and biometric systems may cause anxiety. The ensuing effects can only be judged by experts, so clearly more technological expertise is required.

With advancing digital technology, optimal privacy and confidence will remain a mathematical limit. However, the situation remains the same and so we have to proceed with concrete and critical action and with the help of a comprehensive approach. In this context, the report by the American Federal Trade Commission *Protecting Consumer Privacy in an Era of Rapid Change* (March 2012) mentions the technology-oriented Privacy by Design as the first matter of importance, combined with simple options for consumers, and transparency. The traditional privacy approach remains important, but a comprehensive technological focus now has the highest priority.

2 What is privacy?

2.1 A first outline

Assuming that privacy is a fundamental human right, that there are different flavors, that privacy is a matter of human civilization, as some say, and essential to the economy, is it not a downright shame that there is so much fear, uncertainty and doubt at the moment?

This is all the more true for digital privacy and the value of Personally Identifiable Information: in commercial transactions, in health care, for energy management, in the relationship between citizens and authority etc. Making personal and behavioral data available in exchange for efficient tailor-made service provision can engender an excellent deal with institutions, companies and authorities, as long as we know what is happening to our data and what the risks are. If this is known and arranged with a view to the future, we can then make deliberations and agreements and, as it were, take our Vendor Relationship Management (VRM) into our own hands or tender it out.

To some extent, fear, uncertainty and doubt are just part of our nature, as privacy is typical of the fragile individual who has to stand his ground in the vortex of modern society with all the conflicts of interest that are part and parcel of it. In this digital era of more and more Privacy-Invasive Technologies and data surveillance, no efforts must be spared to remove the sting of fear.

This is done by focusing on personal Total Data Management – in other words, control over our PII, our Personally Identifiable Information. In this context, technology is central in the balance – or race, if you like – between Privacy-Invasive Technologies (PITs) and Privacy-Enhancing Technologies (PETs).

Ideally, this balance has to be established in practice through Privacy by Design in a "fully automatic" and extremely meticulous manner. It means that the PETs have to be integrally adjusted and geared to the correct procedures, regulations, physical environment etc., as proposed at the end of section 1.5 and in the conclusion.

For your reference, in this chapter we define the (digital) privacy theme by means of seven different denominators. We conclude with the increasingly important role of Big Data and a game to practice privacy in social networks.

2.2 Privacy is a fundamental human right

> *No-one should be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks on his honour or reputation.*
>
> *Universal Declaration of Human Rights, 1948, section 12*

In the Universal Declaration of Human Rights of the United Nations, privacy is an unalienable human right and is mentioned as such in charters, constitutions, regular laws and treaties throughout the world. Resolution A/HRC/20/L.13 of July 2012 of the United Nations Human Rights Committee – about "promoting, protecting and having human rights on the Internet" states that all human rights need to be protected offline and online, particularly freedom of speech. Moreover, this is conducive and even vital to economic transactions.

PRIVACY
& FREE
SPEECH
IT'S GOOD
FOR BUSINESS

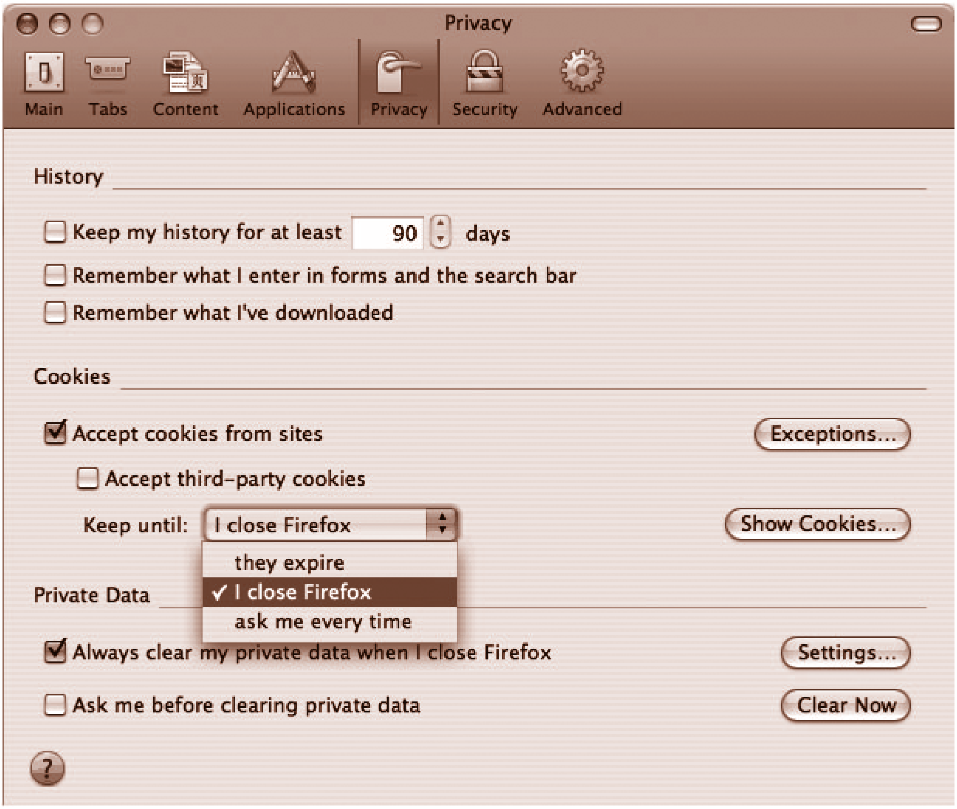
http://www.dotrights.org/business/primer

2.3 Privacy comes in different flavors

The first privacy act dates from 1361, when peeping and eavesdropping were made punishable in England. Modern views with regard to privacy distinguish different categories – such as personal, informational, organizational, spiritual and intellectual

– of “bodily privacy (private parts), territorial privacy (private places), communications privacy (private messages), information privacy.” Our online privacy is usually called *ePrivacy*. Digital privacy is not necessarily online and, according to the letter, informational privacy need not necessarily be digital. The diagram on page 26 gives an indication of what the term “digital Personally Identifiable Information” nowadays means.

Apart from the fact that there are different kinds of privacy, the level of privacy may also differ. An example of this can be seen in our browser settings:



History

☐ Keep my history for at least 90 days

☐ Remember what I enter in forms and the search bar

☐ Remember what I've downloaded

Cookies

☑ Accept cookies from sites

☐ Accept third-party cookies

Keep until: I close Firefox

they expire
✓ I close Firefox
ask me every time

Private Data

☑ Always clear my private data when I close Firefox

☐ Ask me before clearing private data

Levels of privacy are also found in the consumer data collected by organizations (illustration by Magenta Advisory):

1. Identification data

- Name
- Address
- Phone number
- Invoicing information
- Date of birth
- Email address
- IP address

2. Behavioral data

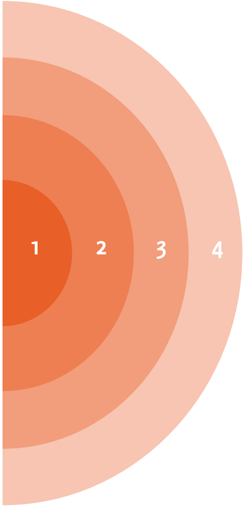
- Purchasing history
- Search and web-browsing history
- Salary information
- Likes on Facebook
- Rating & Reviews

3. Derived data

- Profitability
- Loyalty
- Interest
- Behavioral models
- Analytical models

4. Permission and preferences

- Accepted terms and conditions
- Marketing permissions
- Orders (e.g. newsletter)
- Settings



It is the uncontrolled combination of this kind of digital data that is currently a source of great concern in terms of privacy.

2.4 Privacy is a matter of human civilization

The well-known and even somewhat controversial Russian-American writer Ayn Rand (1905-1982) equated our social civilization concisely with optimal privacy:

> *Civilization is the progress toward a society of privacy. The savage's whole existence is public, ruled by the laws of his tribe. Civilization is the process of setting man free from men.*
>
> *A. Rand (1943),* The Fountainhead

In the eighteenth century, the French political thinker Jean-Jacques Rousseau rather ironically put it as follows:

> *The first man who, having enclosed a piece of ground, bethought himself of saying "This is mine," and found people simple enough to believe him, was the real founder of civil society.*
>
> *Rousseau (1754),* Discourse on the Origin and Basis of Inequality among Men

We need not necessarily agree with the nuances of these observations to appreciate that, from a digital and online point of view, the difference between mine and thine is becoming increasingly obscure nowadays, as is the distinction between public, personal and secret, or that between free of charge and paid for. What does that say of our "civilization"? Are we losing it; did not social civilization always have an adverse effect; should we move with our times and stop moaning about how human constructs such as privacy are shifting the way they do?

2.5 Privacy is essential to the economy

The article "Privacy: Its Origin, Function, and Future" (1979) in which the American economist and professor Jack Hirschleifer emphasizes the economic dimension of privacy, starts with Rousseau's view. The economic dimension explicitly manifests itself in the influential *Framework for Global Electronic Commerce* (1997) by the Clinton administration, as we have seen in section 1.2.

Privacy, Hirschleifer said in 1979, is nowadays not so much a traditional matter of "secrecy" – of withdrawal and of keeping things under cover. It is rather the "autonomy within society" that is central. This autonomy of individuals and groups is synonymous with active economic action. The way in which this could be related to uncertainty and information was something Hirschleifer was specialized in: what does it mean when people do not really know and cannot assess what is known about them? Privacy was "a way of organizing society" rather than of "withdrawal," as Hirschleifer literally underlined it in his article.

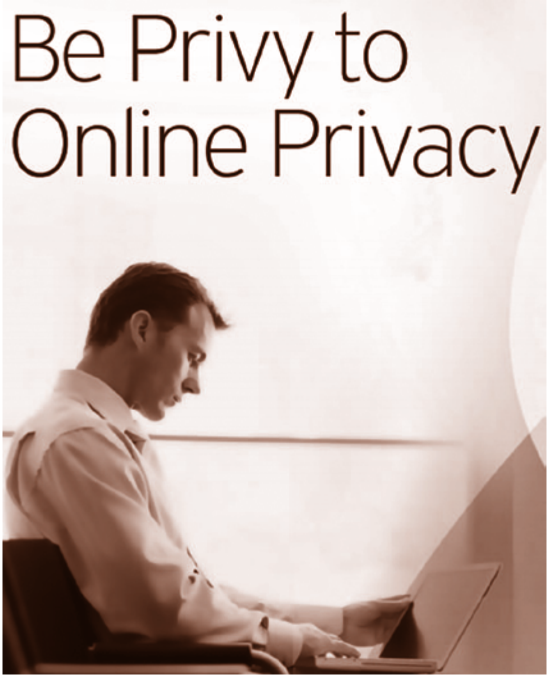
2.6 Privacy is personal Total Data Management

According to the writer Gabriel García Márquez, each individual has three kinds of lives: a public life, a private life and a secret life. As early as 1948, George Orwell described in his book *1984* what devices and the Internet could do in this context:

> *It was terribly dangerous to let your thoughts wander when you were in any public place or within range of a telescreen. The smallest thing could give you away.*

In those days that was a gross exaggeration and it still is in our time, fortunately, but it certainly reflects the fear with which we experience the present *surveillance & dataveillance society*. In the street and online, all conceivable dataflows can be monitored on a permanent basis.

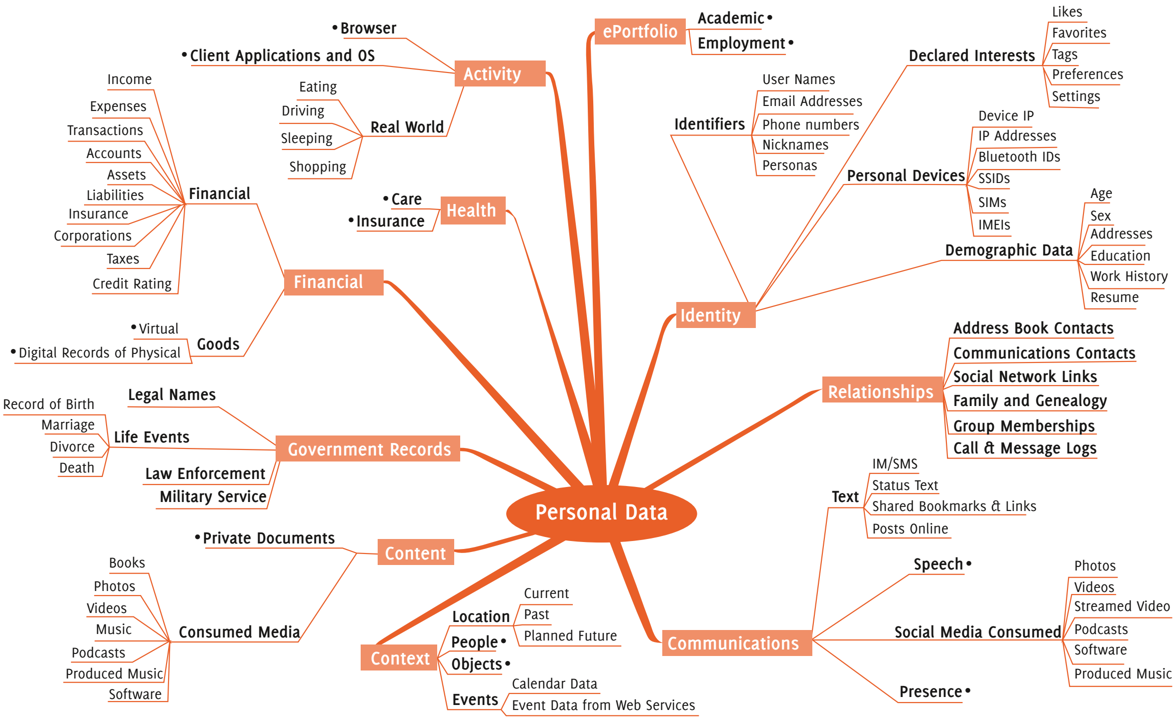
It seems that the only place where we have some privacy is in the toilet at home. It is not without reason that *privy* is related to *privacy* and *private*. The title of this *Digital Life eGuide* plays with that relationship in meaning:



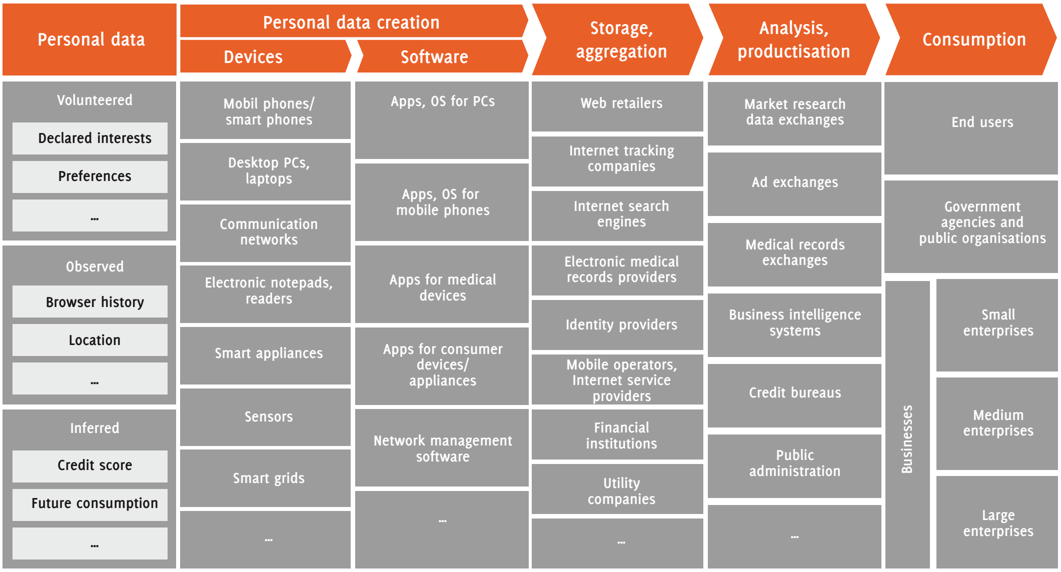
The difference between public, private and secret is the essence of the privacy theme, not in the least in the context of private data and other personal information. Their protection – data protection, *Datenschutz* – is covered by law.

The current European *Data Protection Directive* will be changed into a binding law for all member-states, and is meant to become effective as of 2015. With all the digital activity that we have today, the distinction between public, private and secret is more fluid and fuzzier than ever.

Kaliya Hamlin, also known online as Identity Woman, made the mindmap below showing the cloud with personal digital data or Personally Identifiable Information (PII) that is hanging around us all to a greater or lesser extent: partly public, partly private and partly secret. All in all, this provides a complete picture at any time of who we are, what we think, and what we find interesting; in other words, what we might be ready to pay for or what we might be blackmailed with one way or another.



In diagram form, according to the World Economic Forum report *Personal Data: The Emergence of a New Asset Class*, the PII life cycle from creation to consumption looks as follows:



Personal data
- Volunteered: Declared interests; Preferences; ...
- Observed: Browser history; Location; ...
- Inferred: Credit score; Future consumption; ...

Personal data creation

Devices
- Mobil phones/ smart phones
- Desktop PCs, laptops
- Communication networks
- Electronic notepads, readers
- Smart appliances
- Sensors
- Smart grids
- ...

Software
- Apps, OS for PCs
- Apps, OS for mobile phones
- Apps for medical devices
- Apps for consumer devices/ appliances
- Network management software
- ...

Storage, aggregation
- Web retailers
- Internet tracking companies
- Internet search engines
- Electronic medical records providers
- Identity providers
- Mobile operators, Internet service providers
- Financial institutions
- Utility companies
- ...

Analysis, productisation
- Market research data exchanges
- Ad exchanges
- Medical records exchanges
- Business intelligence systems
- Credit bureaus
- Public administration
- ...

Consumption
- End users
- Government agencies and public organisations
- Businesses: Small enterprises; Medium enterprises; Large enterprises

Personal digital data in all shapes and sizes collectively form the domain of digital privacy. What matters then is not that our valuable PII remains secret at all costs, but rather that we can control which information we are prepared or are not prepared to exchange or sell, as Jack Hirschleifer suggested with his *autonomy* in 1979. To ensure an optimal enforcement of that autonomy as an organizing and economic principle, we need to be constantly aware of the way our PII relates specifically to the two overviews above, of what "leaks away" unintentionally, and of how that information is used.

2.7 Privacy is a matter of trade-offs

When we say that privacy – or simply feeling free or good – is essential to a well-oiled digital economy, the economic concept of *trade-off* immediately comes to mind. Situation-wise and individually we make different choices as to what we will or will not be prepared to allow when it comes to collecting, sharing and using information. After all:

- *Privacy is "the subjective condition that people experience when they have power to control information about themselves and when they exercise that power consistent with their interests and values."*
- *There is no free lunch: We cannot escape the trade-off between locking down information and the many benefits for consumers of the free flow of information.*

Berin Szoka, Senior Fellow, The Progress & Freedom Foundation, 7 December 2009

Privacy is an object of exchange on many fronts: *trade-off is the name of the game.* When parents are nosing in their children's Facebook posts, there is a trade-off between privacy and upbringing. In our digital age, we are even referring to the *privacy paradox:* the wish, on the one hand, to remain anonymous, and the practically unbridled urge to share one's deepest feelings with the world on the other.

Another well-known example is the trade-off between privacy and health. An Electronic Health Record may encroach on our privacy, but we benefit from it in terms of expectancy and quality of life. The same is true of the most current cookie analyses on the Internet and a better service by organizations to customers and prospects. There are various kinds of privacy trade-offs, for instance:

- privacy versus upbringing
- privacy versus health
- privacy versus the fight against fraud
- privacy versus better service

- privacy versus efficient energy systems
- privacy versus self-expression
- privacy versus security.

As (digital) privacy is a trade-off, it is by definition an economic commodity. Faithful to the upright tradition of Hirschleifer, this is also argued by Alessandro Acquisti, co-director of the Carnegie Mellon Center for Behavioral Decision Research, in the paper *The Economics of Privacy*. The economy that is increasingly developing around privacy goes from *mining* and selling personal information to the purchase of products aiming to protect our privacy as consumers.

One of the main trade-offs is privacy versus security, in the sense of being able to move in the physical and digital space without seeing one's physical and ethical integrity challenged. We find it acceptable when the government checks the Internet searching for child porn, we agree to camera surveillance and having our fingerprints taken, etc. But at the same time, skepticism and fear are growing. Nowadays, the Big Brother sentiment is at odds with the opportunities offered by Big Data, in both a commercial and a social sense.

Long before data volume, variety and velocity were topics, Big Brother was already an issue. The history of the records, the censuses, all sorts of recordings and later the population statistics are closely connected with this skepticism, which is often government-oriented. In addition, digital and media go hand in hand nowadays, which has its most appreciable effect for the average citizen in the means used by our surveillance society. For example, this theme was extensively elucidated in *A Report on the Surveillance Society for the [British] Information Commissioner* (2006) by the Surveillance Studies Network.

Lively debates are going on about the trade-off of security versus privacy and the necessity of a trade-off is even contested by people such as Daniel Solove, for example, in his book entitled *Nothing to Hide: The False Trade-off between Privacy and Security* (2011). The argument of the VRM camp is that, thanks to Vendor Relationship Management, privacy need not be a trade-off at all.

In the American Civil War of 1861-1865, Big Brother sentiment received a powerful boost when the population register was used to locate possible military camps in the Southern states. The rise of totalitarian regimes and ideologies, and their disastrous effects, inspired science fiction authors to create scenarios that are so dystopic that they dispel all inclination for any Big Data-like society whatsoever. We are familiar with the classics of the genre:

- **1924** *We*, Yevgeny Zamyatin
- **1932** *Brave New World*, Aldous Huxley
- **1948** *1984*, George Orwell
- **1951** *Foundations*, Isaac Asimov
- **1951** *Fahrenheit 451*, Ray Bradbury

Each of these books lets us have a specific look at how the individual's behavior can be watched and monitored with the help of technology. Later, at the time of the Cold War, America had far more confidence in the government and people were more afraid of the enemy than of the Big Brother in their own country who was spying on them. Senator McCarthy's hunt for communists in the fifties, for example, had a broad support basis among the population.

2.8 Privacy is fear, uncertainty and doubt

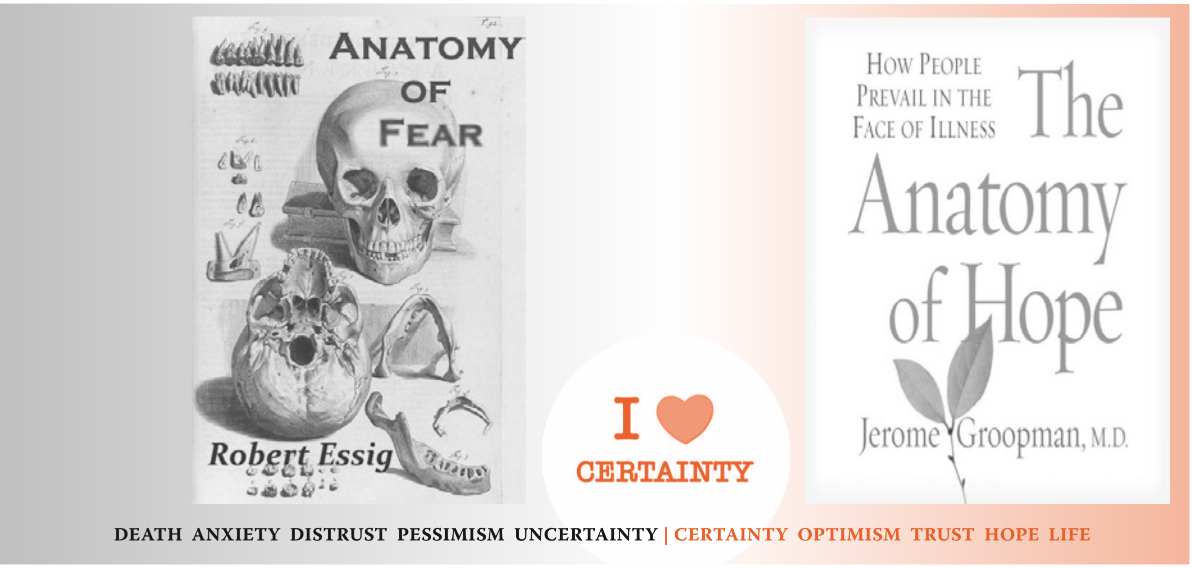
As early as 1966, William Douglas, the longest serving member of the Supreme Court of the United States, said the following about uncertainty with regard to technology-related privacy:

> *We are rapidly entering the age of no privacy, where everyone is open to surveillance at all times; where there are no secrets from government.*

Nowadays living in a surveillance and dataveillance society is considered acceptable. On the one hand, we have to cope with PITS (Privacy-Invasive Technologies) and, on the other, with PETS (Privacy-Enhancing Technologies). The American National Security Agency is currently building a quantum supercomputer, named Vesuvius, to enable constant monitoring of digital data flows of literally everything and everyone in the world. In the name of national security and protection of the democracy, and in all secrecy of course.

Bob Englehart in the Hartford Courant, *22 January 2006, http://www.gsmnation.com/blog/2012/09/25/the-fbi-wants-permission-to-wiretap-your-facebook-account/google-and-the-feds/*

This secrecy also covers the use of Palantir investigation technology, for example, and the NSA's relationship with Google, among others. This extends far beyond *Database Nation: The Death of Privacy in the 21st Century* by Simson Garfinkel (2001). The spectrum of *Fear, Uncertainty & Doubt*, covering fragile individuals and minority groups, can be represented as follows:


ANATOMY OF FEAR
Robert Essig
I ♥ CERTAINTY
HOW PEOPLE PREVAIL IN THE FACE OF ILLNESS
The Anatomy of Hope
Jerome Groopman, M.D.
DEATH ANXIETY DISTRUST PESSIMISM UNCERTAINTY | CERTAINTY OPTIMISM TRUST HOPE LIFE

We see five categories to the left and five to the right of the thin blue line that marks our fragile sense of privacy. Directly linked to the central yearning for security, which is closely connected with the familiar trio of *data – information – knowledge/understanding*, is the methodical doubt expressed by Descartes in the seventeenth century. This *doubt* may all too easily turn into fear if we fail to dispel it. As fear has a paralyzing effect, we wish to deal with it adequately by trying to analyze it, to explain its components, to define its anatomy, so that we can neutralize it.

In this context, it is remarkable that an anatomy of hope, the opposite of fear, is also often described from a negative perception; in the illustration above, it is from the standpoint of illness. And indeed, we often regard our privacy as ill, or at any rate we feel that it is developing in an unhealthy direction.

As far as our sense of privacy is concerned, we have extremely negative feelings most of the time, being afraid of Big and Little Brothers who are threatening and overshadowing us, particularly from a technological point of view. It has been repeatedly stated of late that the best thing to do is forget all about privacy as a privilege in this digital age.

It is this very aspect of "forgetting" that has become a big issue thanks to the Internet, all databases in existence, and also Big Data. *The Right to Be Left Alone* has been further refined into *The Right to Be Forgotten* (2012) by EU commissioner Viviane Reding. But one may well wonder if this can be guaranteed and, if so, through which technologies and procedures. With Abine's DeleteMe app? There is an enormous danger that we will eventually end up in the red zone, with all the economic and social consequences. In this way, privacy may well turn into a showstopper for the wonderful opportunities that datamining and Big Data have to offer in many fields of interpersonal relations.

The Chaos Computer Club raises the alarm

In this context, managing Big Data clumsily is one side of the matter, while deliberately breaking rules is another. These extremes can be illustrated with two examples from Germany. At the annual conference of the Chaos Computer Club in late 2011, it became clear that the smart energy meters of the supplier, Discovergy, were poorly secured. Energy consumption was measured every two seconds for no apparent reason and, in addition, the dataflows were not encrypted. In every household, the consumption per separate device could thus be accurately recorded, with all the consequential possibilities for an analysis of viewing habits and Internet use, for example. The data obtained by the researchers came from "smart meters" that were sealed by way of standard procedure. The poor security might also have enabled hackers – had they managed to penetrate further into the system – to bring the energy supply of millions of households to a standstill.

Without a doubt, the Vesuvius comprehensive data-monitoring project of the abovementioned secret American National Security Agency puts the matter of digital privacy in a particular light. This is also true of the discovery by the same Chaos Computer Club of a computer virus launched by the German government, also in 2011. The code was capable of completely infiltrating a computer, monitoring all actions, storing them and introducing new viruses. Camera, screenshots, Internet telephone traffic, key strokes and of course all hard disk files were completely under control of the monitoring software, reported the *Frankfurter Allgemeine Zeitung.*

2.9 Privacy and Big Data

Ever since the development of media technology such as photography, telephony and telegraphy in the nineteenth century, privacy has become an increasingly important point of interest. The maxim *Privacy Is the Right to Be Left Alone* originated in the America of the 1890s. This is when the development of technology and our need of privacy were seriously at odds for the first time:

> *Recent inventions and business methods call attention to the next step which must be taken for the protection of the person, and for securing to the individual what Judge Cooley calls the right "to be left alone." [...] Numerous [...] devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the rooftops."*

To date, the validity of this quotation still stands. Without the photography, the newspapers and the word "mechanical," which we have deliberately omitted from the quotation, no one could have suspected that these words date from 1890, and have been taken from the article "The Right to Privacy" by Samuel Warren and Louis Brandeis in the *Harvard Law Review*. In fact, this is what primed the entire modern debate on privacy. And even today, photography and paparazzi occupy center stage when it comes to privacy issues.

That Big Data is an extra cause of concern these days also became evident when Schufa, the largest German credit-rating firm, announced that it intended to link information from Twitter, Facebook and LinkedIn to its 66-million customer database for the sake of better profiles.

This attack on the people's right to have control over their own data gave Germany reason to fear "American circumstances." Ilse Aigner, the Minister with consumer rights in her portfolio, stated that social networks must not be systematically used to assess credit applications.

In 2012, the following three articles, dealing with the combined theme of privacy, data protection and the rise of Big Data, were among those that caught the eye:

- The first one was entitled "Privacy in the Age of Big Data: A Time for Big Decisions," and was published in the February issue of *Stanford Law Review*.
- Number two, "The Challenge of Big Data for Data Protection," first saw the light of day in the May issue of the *Oxford Journal on International Data Privacy Law*.
- And the third, "Privacy by Design in the Age of Big Data," came from the Office of Ann Cavoukian, the Information and Privacy Commissioner of Ontario, in June. Her co-author is IBM's Big Data guru Jeff Jonas.

The article "Big Data for All: Privacy and User Control in the Age of Analytics" on the website of the *Stanford Law Review* magazine (February 2012) gave a good idea of what is going on with privacy and data protection in the light of Big Data. The reasoning is as follows: Big Data is reality; it is extremely valuable, but at the same time it fuels uneasiness about privacy. So a good balance needs to be created between organizations and individuals. At the very beginning of the book, authors Omer Tene and Jules Polonetsky make the following seven points:

The Big Data reality

1. *The amount of information at the disposal of organizations and authorities has expanded, due to developments in data mining and analytics, and the enormous increase in computing power and data storage.*
2. *Raw data can now be analyzed without the help of structured databases. This way, it is much easier to demonstrate interrelationships, while new unthought-of applications for existing information are beginning to emerge.*
3. *At the same time, the growing numbers of people, devices and sensors that are linked by means of digital networks have caused a revolution in creating, communicating, sharing and accessing data.*

The value and privacy challenge of Big Data

1. *Data are of great value to the world economy as the raw material for innovation, productivity, efficiency and growth. At the same time, the flood of data poses privacy issues that may result in regulations that bring the data economy and innovations to a standstill.*
2. *To find a balance, policy makers need to address a number of the most fundamental privacy concepts, such as the definition of Personally Identifiable Information (PII), the way it can be controlled by the individual, and the principles of minimal and effective use of data.*

A good balance for organizations and individuals

1. *When individuals have data at their disposal in an accessible manner, they can share the wealth of the information. On that basis valuable client applications can be developed.*
2. *It is obvious that organizations are obliged to be quite explicit when it comes to their decision criteria, for in a Big Data world it is usually the conclusions that give cause for concern and not the data themselves.*

The article "Big Data for All," to be published in the *Northwestern Journal of Technology and Intellectual Property*, provides an overview of the advantages of Big Data in different fields and economic sectors. Subsequently the drawbacks are discussed, and this is followed by a number of central challenges and finally by a number of solutions.

The concept of Privacy by Design provides a fundamental underpinning: the worrisome effect of Privacy Invasive Technologies must be counterbalanced by an intense combination of Privacy Enhancing Technologies, regulations, policy, procedures and responsible behavior by all parties involved.

In this way Big Data, too, should become a win-win situation for all. Politics, businesses, authorities and individuals all over the world are looking forward to seeing that ambition and promise realized.

Privacy, technology and the law

The idea is that in the years to come there will be a huge change for the better in the relationships between privacy protection, digital technology and regulation. This is essential for the development of the economy and of social relationships. In 2011, the 112th American Congress (the Obama I administration) was the first to install a special Senate Committee with the clear name: *Privacy, Technology and the Law*. The committee has the following five main tasks in its portfolio:

- *Supervision of regulation and policy with regard to the collection, use and dissemination of commercial information by the private sector, including behavioral advertising, privacy in social networks and other online privacy issues.*
- *Enforcement and implementation of regulation and policy with regard to the privacy of commercial information.*
- *Use of technology by the private sector to protect privacy, enhance transparency and stimulate innovation.*
- *Privacy standards for the collection, storage and management, use and dissemination of commercial Personally Identifiable Information (PII).*
- *Privacy implications of new or emerging technologies.*

In our Big Data business reality, we are now heading on a worldwide scale toward a fundamental emphasis on transparency and choice, toward "informed consent" and clear "opting out" possibilities for individuals. In this context a balance between PITS and PETS, combined with clear regulations and procedures through Privacy by Design, seems to be the best and most comprehensive solution. This subject is dealt with in Chapter 3.

2.10 A game to practice privacy

Cultivating one's awareness and individual responsibility is also part of Privacy by Design. For a short time now, a privacy game has been available on http://www.open.edu/openlearn/privacy to practice the new standards and values on social networks.

Via “Secret or sharing? Play our Privacy Game” you can decide which information you are willing to share and which you had better keep to yourself. The game offers the opportunity to make a small, valuable and perfectly safe bet with your personal data. Via OpenLearn you are playing the computer alone, but you can also challenge your Facebook friends to a multiplayer version of the privacy game, which still has a closed character.

Count not him among your friends who will retail your privacies to the world.
Publilius Syrus, ca 50 BC

3 Privacy by design and the balance between PITS and PETS

(Privacy-invasive versus privacy-enhancing technologies)

3.1 A new taxonomy of privacy

In January 2006, the *University of Pennsylvania Law Review* published an 84-page article by Daniel Solove, currently a professor at George Washington University Law School. In the article with the same succinct title, to which another 25 experts contributed, Solove presented a new *Taxonomy of Privacy*, linked to technology and information.

Digital innovations have become more and more prevalent, but the abstract legal concept of privacy was and still is insufficiently geared to this situation, to use an understatement. Solove c.s. are hard as nails in their assessment:

> ***Privacy is a concept in disarray. Nobody can articulate what it means.*** *[...] Privacy is far too vague a concept to guide adjudication and lawmaking, as abstract incantations of the importance of "privacy" do not fare well when pitted against more concretely stated countervailing interests. [...] This Article develops a taxonomy to identify privacy problems in a comprehensive and concrete manner.*

The notion of privacy should be fleshed out, not least to ensure its unequivocal character in the context of legislation. By early 2006 we had already made considerable progress in the digital age, but concrete new privacy issues were hardly addressed adequately. It was high time to create a comprehensive and clear regulation that would primarily deal with the activities of individuals, organizations and authorities:

> *Technology is involved in various privacy problems, as it facilitates the gathering, processing, and dissemination of information. Privacy problems, however, are caused not by technology alone, but primarily through activities of people, businesses, and the government. The way to address privacy problems is to regulate these activities.*

From this interesting observation of 2006, we have now – seven years later – come to a point where interest in activities is gradually being combined with the development and enforcement of a good balance: between Privacy-Invasive Technologies (PITS) on the one hand and Privacy-Enhancing Technologies (PETS) on the other. This fundamental and integral approach is known as Privacy by Design. The following taxonomy of privacy by Solove et al., dating from 2006, which displays a clear focus on technology and information, acts as a sounding board in that context:

Information Collection

- Surveillance
- Interrogation

Information Processing

- Aggregation
- Identification
- Insecurity
- Secondary Use
- Exclusion

Information Dissemination

- Breach of Confidentiality
- Disclosure
- Exposure
- Increased Accessibility
- Blackmail
- Appropriation
- Distortion

Invasion

- Intrusion
- Decisional Interference

3.2 Personally Identifiable Information and PETS

For a variety of reasons and in a variety of ways, organizations have the Personally Identifiable Information (PII) of employees, customers and other parties at their disposal. In this context, the rules for privacy and data protection must be upheld. Well-designed and well-implemented Privacy-Enhancing Technologies (PETS) are the opposite of Privacy-Invasive Technologies (PITS). They aim to realize the required protection in combination with regulations, guidelines, processes, training etc.

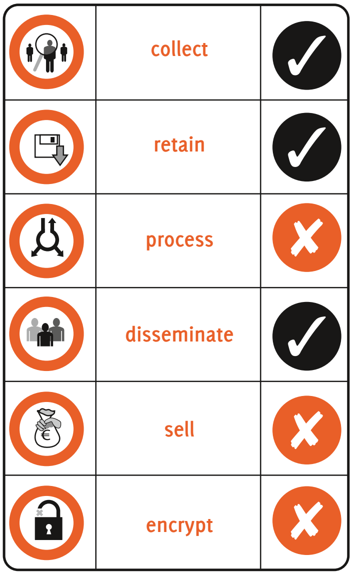
Ideally, PETS have a clear connection with what privacy rules require and intend. Therefore the British Information Commissioner's Office describes PET as:

> *any technologies that protect or enhance an individual's privacy, including facilitating access to their rights under the Data Protection Act.*

In addition, the European Union emphasizes the role of PETS in the designing of information and communication systems, in such a way that any regulation from the perspective of technology is given a firm basis:

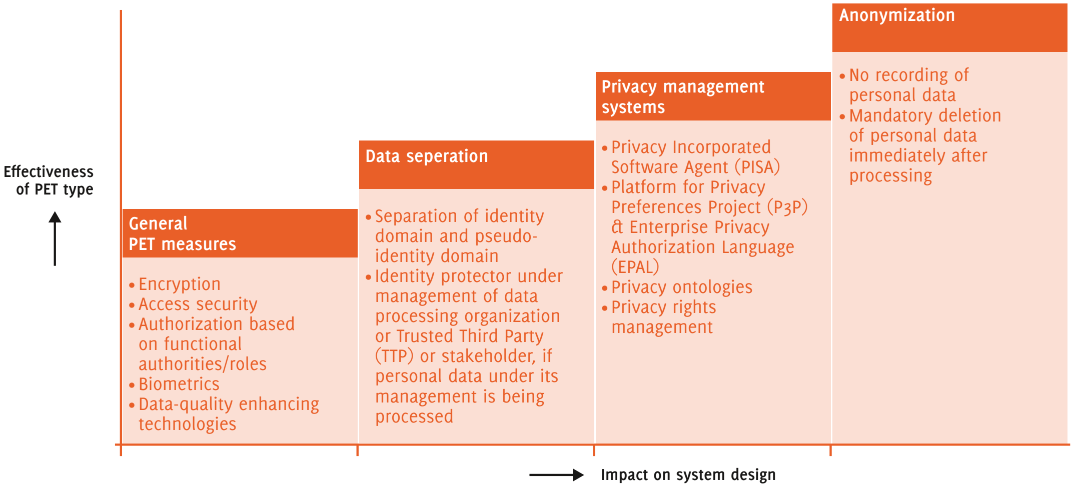
> *The use of PETS can help design information and communication systems and services in a way that minimises the collection and use of personal data and facilitates compliance with data protection rules making breaches more difficult and/or helping to detect them.*

The above-mentioned *Taxonomy of Privacy* by Daniel Solove is a perfectly adequate vehicle for a *Privacy Impact Assessment* framework (see section 4.7) for PETS that intend to prevent privacy-related damage. What is at issue here is what is known as *Fair Information Practices*, the foundation of a digital economy that is worthy of confidence, not least when Big Data are used. Our personal data, our PII or simply our contextual ID come into play here. While thanking Alexander Alvaro, vice-president of the European Parliament, we have reduced the description of his PII pictograms to the following practical set of FIPS *(Fair Information Practice Principles)* on the left.



	collect	✓
	retain	✓
	process	✗
	disseminate	✓
	sell	✗
	encrypt	✗

A concrete overview of PETs is provided by the relevant wiki of the Center for Information and Society: http://cyberlaw.stanford.edu/wiki/index.php/PET. The graph below by Koorn and Ter Hart (2011) provides an overview of the effectiveness of different PET types, compared to the impact on the system design.



General PET measures

- Encryption
- Access security
- Authorization based on functional authorities/roles
- Biometrics
- Data-quality enhancing technologies

Data seperation

- Separation of identity domain and pseudo-identity domain
- Identity protector under management of data processing organization or Trusted Third Party (TTP) or stakeholder, if personal data under its management is being processed

Privacy management systems

- Privacy Incorporated Software Agent (PISA)
- Platform for Privacy Preferences Project (P3P) & Enterprise Privacy Authorization Language (EPAL)
- Privacy ontologies
- Privacy rights management

Anonymization

- No recording of personal data
- Mandatory deletion of personal data immediately after processing

Technology is an indispensable aid, but its success invariably depends on implementation – see Koorn and Ter Hart (2011) for an overview – and adoption. In *Privacy Enhancing Technologies: A Review* (2011), Yun Shen and Siani Pearson of HP Laboratories recommend a focus on the following fields:

- Usability
- Privacy by Design
- Economics of Privacy

As for the last, we usually do not feel that the costs of a considered privacy choice, however small, are worth it nowadays. In practice, the so-called *Willingness to Pay* clearly loses out to the *Willingness to Accept*. A good example is unquestioningly accepting page-long terms and conditions online.

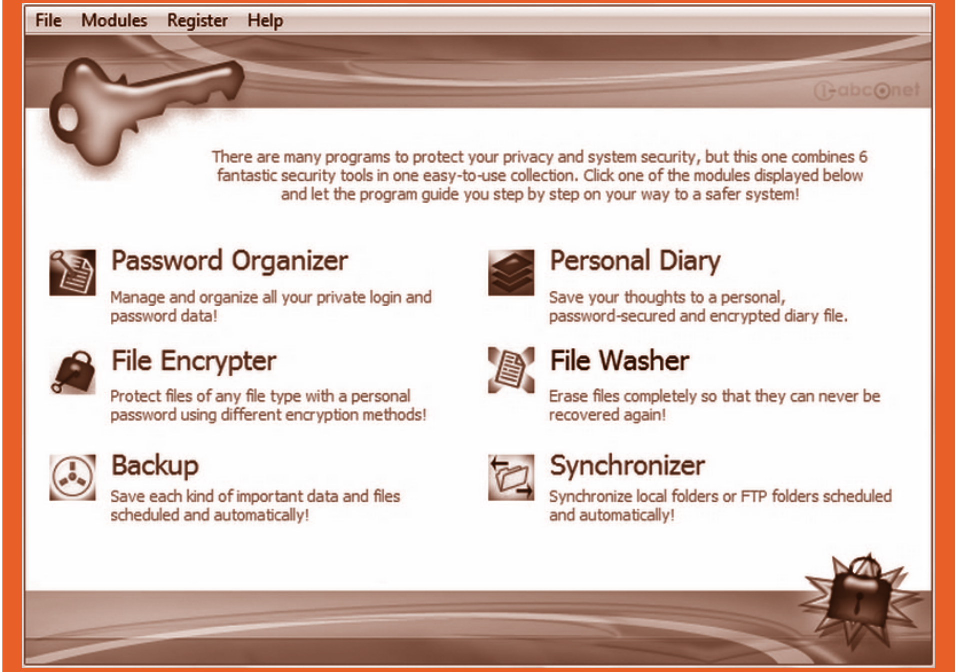
Differential Privacy

In the context of the still extremely relevant theme of database privacy, *Differential Privacy*, which is relatively unknown, needs to be added. It is very difficult to guarantee the individual protection of privacy in databases, even if PII has been minimized. It often turns out that, with a lot of trouble, it is possible to use data from other databases by way of supplement, thus converting the information to the individual level. Differential Privacy neutralizes this re-identification problem by adding white noise, among other things, to the otherwise correct database matter. The quality of the aggregated results is not at stake because of a Differential Privacy approach.

3.3 Privacy according to TNO and TILT

In December 2011, the Dutch organization for Applied Physics Research TNO and the Tilburg Institute for Law, Technology and Society TILT published the report entitled *Trusted Technology: a study into the application conditions for Privacy by Design in the electronic services of the government*. The concept of Privacy by Design is explicitly based on Privacy-Enhancing Technologies (PETS). See, for example, the 350-page *Handbook of Privacy and Privacy-Enhancing Technologies* (2003). That publication is devoted to intelligent software agents.

The PET spectrum has a huge range. This security tool, for aspects such as protected passwords, file encrypter, an encrypted diary file and an option to erase files completely, is a basic PET application at computer level:



There are many programs to protect your privacy and system security, but this one combines 6 fantastic security tools in one easy-to-use collection. Click one of the modules displayed below and let the program guide you step by step on your way to a safer system!

Password Organizer
Manage and organize all your private login and password data!

Personal Diary
Save your thoughts to a personal, password-secured and encrypted diary file.

File Encrypter
Protect files of any file type with a personal password using different encryption methods!

File Washer
Erase files completely so that they can never be recovered again!

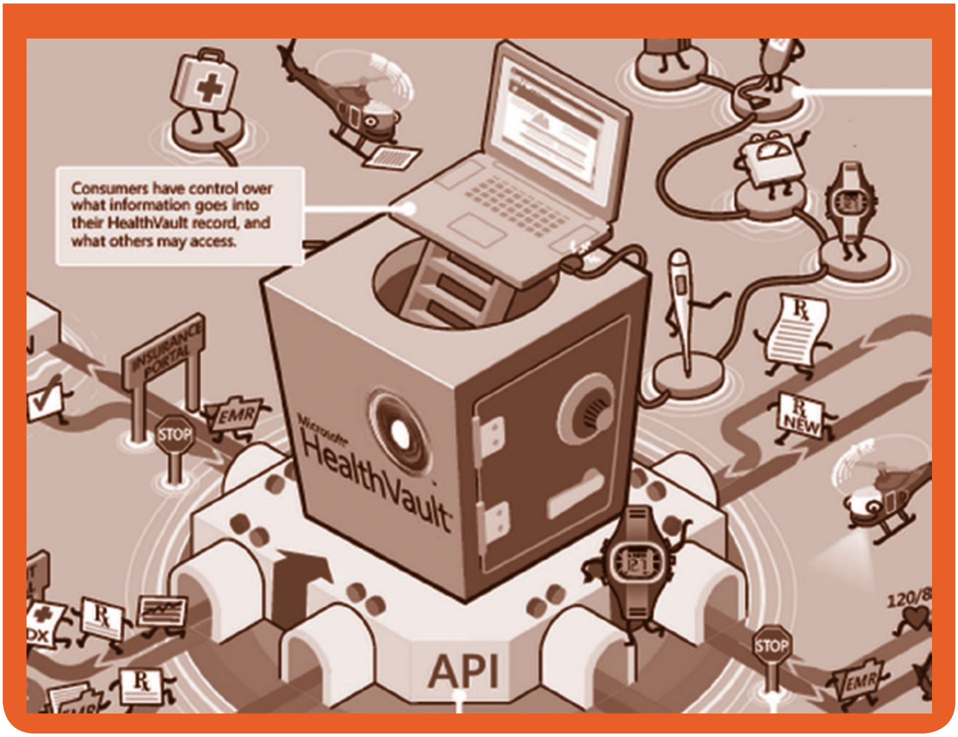
Backup
Save each kind of important data and files scheduled and automatically!

Synchronizer
Synchronize local folders or FTP folders scheduled and automatically!

An overview of the IT-specific *Top Ten Big Data Security and Privacy Challenges* is provided in the report of the same name by the Cloud Security Alliance – including the Fujitsu and HP Labs – of November 2012.

Another focus is that of the Digital Vault. This kind of vault exists in different sorts and sizes, from simple consumer applications – by British Telecom, for example – to patented vault technology by suppliers such as CyberArk, which has been especially developed to come up to the level of the best secured bank vaults.

A concrete application, in line with the new Electronic Health Record, is Microsoft HealthVault. In this type of personal vault individuals can store their health-related information, update it, link it to devices and share it with medical professionals, the drugstore, insurers and other parties without any problem.


Consumers have control over what information goes into their HealthVault record, and what others may access.

Privacy by Design is the use of technical and organizational measures in information systems to avoid invasions of people's personal privacy. If information systems are inherently privacy-friendly, this considerably adds to a sustainable information society.

But, by and large, the report by TNO and TILT explains, the protection of privacy encompasses all activities and measures aimed at the regulation of access to the individual in a situational, relational and informational sense. So this extends beyond the protection of personal data or, in other words, data protection.

At the end of the day, the protection of privacy aims to protect or enhance people's personal autonomy and to reduce their vulnerability to material damage, discrimination and stigmatization, for example, as much as possible. Moreover, privacy is not only meant to protect individuals. The values on which privacy is based also have important social and libertarian dimensions.

Privacy enables people to arrive at individual views and preferences without outside interference. In this way, privacy adds to the multiformity and creativity of society and to the protection and enforcement of the democratic constitutional state. All this according to TNO and TILT.

3.4 E-privacy-related challenges

The view of TNO and TILT is a legitimate, albeit very idealistic one of the e-privacy discussion. A more concrete emphasis on *Trustworthy Social eCommerce* is found on Eprivacy.com, expressed by Philippe Coueignoux. In his analysis “ePrivacy, What’s at Stake?” Coueignoux explains that IT and Internet-related internal fraud, external fraud and explicit privacy issues all cause breaches of confidence that directly put a spanner in the works, as he puts it, of *Economic Activity & Individual Business Valuation*. Coueignoux provides the following useful classification of five e-privacy-related themes that he observes among *Liabilities and Vulnerabilities in the Information Age*.

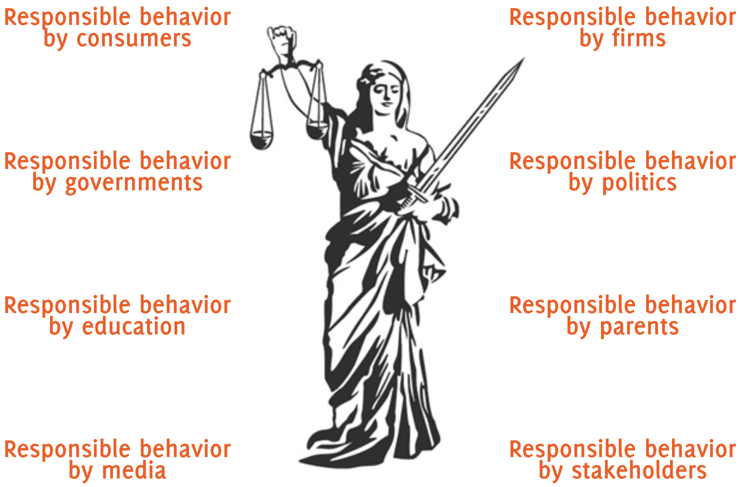
Overview of online privacy issues

1. *Identity:* identity theft, credit fraud, ambush marketing
2. *Ownership:* medical records, marketing campaigns, international data & Safe Harbor (see section 4.6), surveillance, viral marketing
3. *Location:* searching and matching data
4. *Defense (the good side):* protecting, storing, using, distributing and recommending digital information
5. *Offense (the dark side):*
   - stealing time from receivers: spamming, manipulating search results
   - blocking access: denial of service, censorship
   - falsifying the context: copying, plagiarism and forgery, advertising fraud

A recent suggestion by Coueignoux to presidents Barack Obama and Herman van Rompuy is to tax the economic use of privacy; it should be done progressively and separately from the continued enforcement of e-privacy legislation. Coueignoux’s partly technological Privacy & Security by Design solution dovetails with the economic value that personal data has in the age of the Internet.

3.5 Responsible behavior as core value

Responsible behavior is the core value of all ethical-legal themes. On account of its fluidity, this is certainly true of the protection of personal digital data on the Internet. A non-limitative, generic list of kinds of behavior displayed by various stakeholders, with legislation and regulation, jurisdiction and jurisprudence in the centre – preferably internationally harmonized or uniformized – is shown next:


Responsible behavior by consumers
Responsible behavior by firms
Responsible behavior by governments
Responsible behavior by politics
Responsible behavior by education
Responsible behavior by parents
Responsible behavior by media
Responsible behavior by stakeholders

All the possibilities and means at the disposal of Justice for the purpose of its legislative, executive/enforcing and judiciary powers are truly impressive, but we still have a long way to go before it is all realized in actual practice. Ideally, Privacy by Design and Security by Design are at the basis of all efforts in the context of privacy and security.

First and foremost, Privacy & Security by Design involves building in privacy protection by means of Privacy-Enhancing Technologies (PETS), beginning with the system design. Apart from the technical micro-level, the principle should also be effective on an organizational meso-level and legal macro-level. The aim of Privacy & Security by Design is twofold: secure privacy-friendly system designs, and a sustainable information society in day-to-day practice.
More information can be found in the recent report entitled *Operationalizing Privacy by Design: A Guide to Implementing Strong Privacy Practices* by Ann Cavoukian, the Canadian Information and Privacy Commissioner, which is also discussed in the conclusion.

4 Legislation in a state of flux

4.1 We ourselves are the enemies of privacy

The American Supreme Court Judge Alex Kozinski finds it hard to understand that we tend to blame others for our loss of privacy. After all, why do that if we refuse to take our own privacy seriously? Today's exhibitionistic behavior is making it increasingly difficult for American judges to stand up for privacy. It is day-to-day practice, in conjunction with other factors, that decides how the government and the judicature deal with it and, if that practice implies exhibitionism, then that change of standards is a fact.

In a Twitter world where the police can ping us on a smartphone in the wink of an eye, an increasing number of people regard this simply as a matter of having more followers. This was Kozinski's tongue-in-cheek observation at the Stanford Law Enforcement Symposium 2012 on "privacy and its conflicting values":

> *The idea that law enforcement can now ping your cell phone and find out exactly where you are at any time, with no probable cause and no judicial supervision, is greeted with a big collective yawn. In a Twitter world where people clamor for attention, having the police know your whereabouts just increases your fan base.*

4.2 Throwing away my privacy for 50 cents

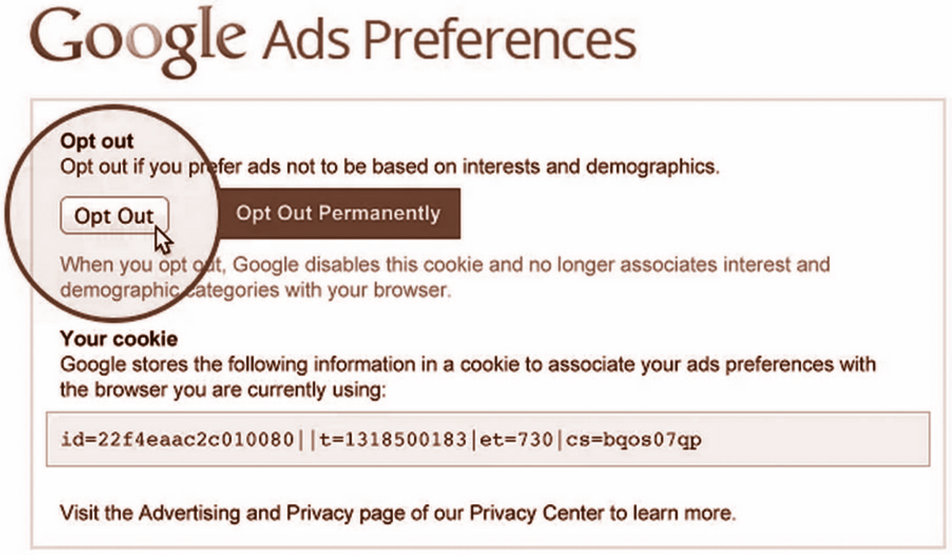
What is our privacy worth to us? A fifty-cent discount on a 7.50 cinema ticket is an excellent online exchange for our telephone number or e-mail address. This appears from the *Study on Monetizing Privacy* by ENISA, the European Network and Information Security Agency of early 2012.

Evidently we do not care much about divulging personal information. Three-quarters of Europeans increasingly regard it as a fact of life. Over 40% of European Internet users say they are asked to supply more data online than is strictly necessary, but just do it all the same. This was revealed by the study *Attitudes on Data Protection and Electronic Identity in the European Union*, carried out in late 2010.

In the following year, Facebook doubled its revenue from advertising to almost 4 billion dollars. This concerns first, second and third-party cookies, which keep a close track of our online behavior and make capital out of it. Consciously supplying information is one thing, but being followed without being fully aware of it is a much larger issue in the personal data concern. For this reason, a stringent opt-in cookie law was introduced in the Netherlands in June 2012, which was liberalized again just before the end of the year in accordance with the intended European standard. For, apart from privacy, entrepreneurship and innovation are also considered to be of paramount importance, not least in economically hard times.

More and more people are saying that we should just forget all about our privacy; after all, something like privacy simply no longer exists in this age of the Internet, just as little as "intellectual property," for example. Perhaps it does not really matter either way because, analogous to this, most companies are scared to death of the reputational damage that might be caused if it turns out that customers are being closely scrutinized without their knowledge, just to optimize the companies' profits.

Large-scale abuse of data on the basis of a few simple cookies is grossly exaggerated and a rigid opt-in rule would be disastrous for entrepreneurship. The prevailing view is that we have to deal with articulate consumers nowadays and, if they are properly informed and the possibilities to opt out are pointed out to them, that is the most attractive economic situation for all parties.



Google Ads Preferences

Opt out

Opt out if you prefer ads not to be based on interests and demographics.

When you opt out, Google disables this cookie and no longer associates interest and demographic categories with your browser.

Your cookie

Google stores the following information in a cookie to associate your ads preferences with the browser you are currently using:

```
id=22f4eaac2c010080||t=1318500183|et=730|cs=bqos07qp
```

Visit the Advertising and Privacy page of our Privacy Center to learn more.

4.3 Privacy no longer the social standard

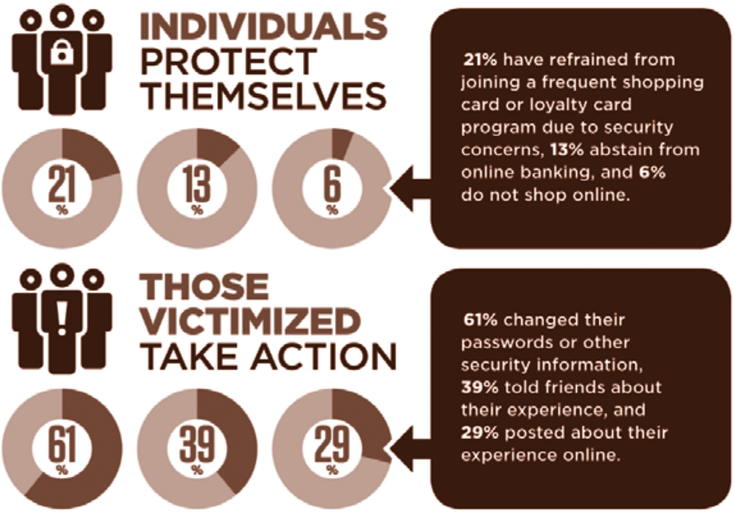
The first one to repudiate the relevance of the digital privacy discussion was Scott McNealy, the then CEO of Sun Microsystems. "You already have zero privacy – get over it," he said in 1999 when Jini was introduced: software intended to link a large number of different devices. For example – as is perfectly normal today – making a photograph that is uploaded automatically to a newspaper via the Internet, so that it ends up in newspaper stands all over the world the same day.

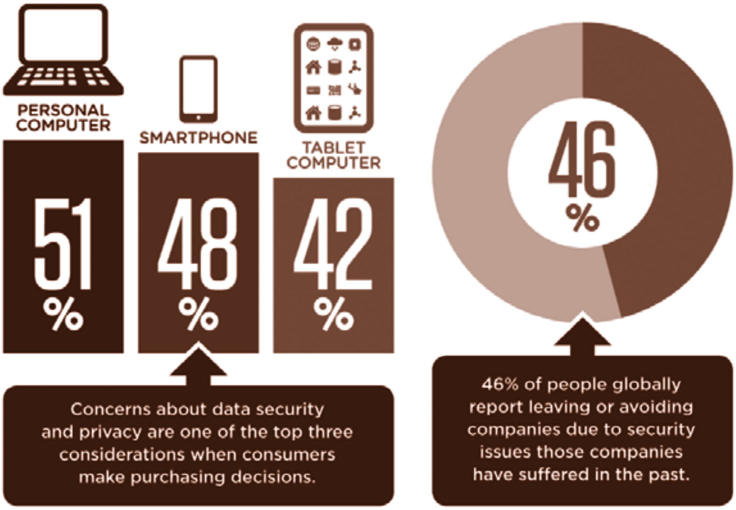
One decade later, in early 2010, Mark Zuckerberg, the CEO of Facebook, took a similar stand in an interview with TechCrunch, by arguing that privacy was no longer the social standard. "When we started Facebook in my room at Harvard seven years ago," said Zuckerberg, “we wondered if people would put information online at all. But in a few years’ time the standards entailed in privacy have changed completely. And we are simply moving with the times.”

Kozinski, ENISA, McNealy and Zuckerberg arrive at the same conclusion from different perspectives: in the past decades, and years even, the views of privacy have undergone a tremendous change, which is why, apart from standards and values, the rules and regulations are likewise moving with the spirit of the times.

4.4 Privacy & security: the new drivers of brand, reputation and action

The latter is perfectly obvious from Edelman’s info diagram (shown below), which has been drawn up on the basis of a survey in 2012 among 4050 individuals from 7 different countries. Edelman calls privacy and security the “New drivers of the brand.” Privacy must become a core task for every organization. Almost half of all the consumers interviewed state they tend to avoid companies that have failed to protect data properly. The fact that security breaches, for example, may go viral in an instant and generate a tremendous amount of negative publicity is one of the reasons for organizations to engage with these new competencies with great enthusiasm.





If much commotion is made about a privacy breach at all, the organization in question will usually have a quick and easy statement prepared that it considers to be in line with prevailing regulation and standards. By and large, privacy matters are considered very seriously in economic activities. But – on account of its high-profile Big Brother issues – privacy is a field of increasing controversy and (hyper)sensitivity, and there are many people who are all too eager to sound the alarm.

There is no denying that weak security of systems, far-reaching powers for authorities, a hodgepodge of obsolete laws and regulations, actions by groups such as WikiLeaks and Anonymous, cybercrime and cyberwarfare etcetera strongly determine the sentiment with regard to e-privacy and data protection. It would be wrong, however, to indiscriminately continue this negative sentiment in the context of the day-to-day privacy practices of organizations engaged in business activities.

In the context of the information society, the Surveillance Society and Big Data, everything that concerns privacy is very much in a process of flux at the moment, while measures are further refined and focused. For example, the idea is that the European Guideline, on which national governments currently base their own privacy laws and regulations, will be replaced by a stringent regulation in 2015; in short, by a universally applicable European law. Such uniformization is advisable in order to create an economic *level playing field*.

However, without a concrete case, the question as to what is and what is not allowed with regard to privacy can hardly be answered satisfactorily. But even if there is a case, there are many comparative assessments and pros and cons in the balance. In addition, it is difficult to explain current legislation while it also displays so many gaps, according to the opinions of various experts we consulted for this book.

4.5 Effectively formulating and rationalizing plans

When one has commercial plans, the following points are vital:

1. They have to be formulated as precisely as possible and, ideally, one should be able to indicate where one would like to be in, say, five years' time.
2. One has to demonstrate on the basis of well-reasoned arguments why the intended actions are fair and socially justified in relation to the focus group of customers, staff, prospects and suspects, for instance.
3. Subsequently it can be decided in consultation whether or not there is an acceptable legal form available. The target group should at least be informed of the plans (obligation to provide information) and there should be no gap between rhetoric and reality.
4. Transparency is of vital importance and anything reminiscent of discrimination and applying double standards, like *dual pricing*, must be avoided.
5. Explicitly asking for permission is only required in the case of serious matters such as health and criminal law, etc.
6. Big Data practices, like gathering personal information from a variety of sources and subsequently drawing conclusions as a result, e.g., segmentation, must be explained carefully.
7. *Straight-through processing* – in general, processing data without human intervention is not allowed.

4.6 Guidelines of the OECD

A practicable and universal first point of departure for the protection of privacy and data is the OECD guidelines of 1980, summarized at http://oecdprivacy.org. These *OECD Privacy Principles*, eight in all, concern respectively:

- *Collection Limitation:* data must not be gathered haphazardly or illegally, and, if applicable, they should be gathered with the knowledge or permission of the party concerned.
- *Data Quality:* the accuracy of the data must be guaranteed.
- *Purpose Specification:* it must be clear what the data will be used for.
- *Use Limitation:* the use of the data must be limited.
- *Security Safeguards:* the security of the data must be guaranteed.
- *Openness:* it must be perfectly clear which data are being collected and what will be done with them.

- *Individual Participation:* during the entire process of gathering and using the personal data etc., the individual must be actively involved and offered easy access, so that the person in question is informed adequately and is in a position to take action.
- *Accountability:* the "data controller" is responsible for the compliance with these eight Fair Information Practice Principles (FIPS).

These principles can be consulted in detail and in their context on the IT Law Wiki (http://itlaw.wikia.com/wiki/The_IT_Law_Wiki).

All in all, this is perfectly in tune with the European view of privacy. The American *Consumer Privacy Bill of Rights* of February 2012 (officially: *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*) contains an attachment in which America's own Fair Information Practice Principles (FIPS) are compared with those of the OECD among others.

A 119-page document of the European Commission of January 2012 presents the so-called "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)." The idea is that this proposal will come into force in 2015.

At the annual European Data Protection and Privacy Conference in 2012, the American ambassador William Kennard said that the EU's new data protection rules threaten the co-operation between the European and American police and judiciary, because hundreds of investigation regulations that are functioning perfectly well would have to be adjusted.

Apart from the above OECD guidelines, the OECDprivacy.org website mentions another three privacy frameworks: the *Asia-Pacific Economic Cooperation (APEC) Privacy Framework*, the *United States Department of Commerce Safe Harbor Privacy Principles*, and the *Generally Accepted Privacy Principles* (GAPP). The worldwide trend with regard to privacy and security regulation is one of harmonization, uniformization and standardization.
The US-EU **Safe Harbor treaty** aims to help American organizations meet the EU rules with regard to the protection of personal data. It includes checklists,

self-certification and a workbook. The following seven Safe Harbor principles are central:

- *Notice* – Individuals must be informed that their data are being collected and about how they will be used.
- *Choice* – Individuals must have the ability to opt out of the collection and onward transfer of the data to third parties.
- *Onward Transfer* – Transfers of data to third parties may only take place to other organizations that follow adequate data protection principles.
- *Security* – Reasonable efforts must be made to prevent loss of collected information.
- *Data Integrity* – Data must be relevant and reliable for the purpose they were collected for.
- *Access* – Individuals must be able to access information held about them, and correct or delete it if it is inaccurate.
- *Enforcement* – There must be effective means of enforcing these rules.

That looks good, but after two negative reviews by the European Union in 2002 and 2004, Galexia expressed the following opinion in 2008:

The growing number of false claims made by organisations regarding the Safe Harbor represent a new and significant privacy risk to consumers.

The OECD rules mentioned above are officially called the *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. The complete text can be found in the documents under the heading "Information security and privacy," a topic in the section "Internet economy" on the website.

4.7 Privacy Impact Assessment (PIA)

To keep the theme of privacy and security practical, a *Privacy Impact Assessment* (PIA) can be made to clarify in advance the risks inherent in the implementation of plans. The British Information Commissioner's Office, for example, has such a *PIA Handbook* and distinguishes the following nine steps:

- identifying interested parties
- initial assessment of privacy risks
- decision as to the extent of PIA
- mapping out privacy risks

- consulting interested parties
- making proposals for acceptance
- moderating or avoiding risks
- checking compliance
- planning a review.

The review examines the actions undertaken as well as the effects, which may result in a more extensive or even new PIA.

5 Seven privacy principles and your next step

Privacy is a concept that dates from the fifteenth century, but as early as the first century BC, Publilius Syrus said that we should not associate with "friends" who shout our private matters from the rooftops. This is an excellent message in this era of social media, which just goes to show that privacy issues have existed since time immemorial, and are closely connected with the individuality that is typical of every rational human being.

Everyone likes to exercise discrimination when it comes to his or her privacy. Traditionally the mine and thine and our private domain are at stake, but digital privacy is a particularly ambivalent matter. If we have been branded by information systems as being less creditworthy, for example, this may have long-term effects, as practice has shown many times. And this gives a nasty flavor to the concept of digital DNA.

Privacy scandals have a long half-life: we tend to nurture our suspicion. Only too often have we lapsed into the same old mistakes, somehow or other. Statistically speaking, it reminds us of the wise men who are unable to answer all the fool's questions or remove his fundamental distrust. Anything that is in the public eye tends to take root, not least in the absence of systematic clarity.

Nowadays, everyone is very much aware that privacy, technology and regulation form the triad that should be able to provide sufficient clarity and certainty to generate a better socio-economic digital world. An integral approach to Privacy by Design will have to create a basis for trust so that we may eventually reap the fruits of a digital economy.

The development of Big Data and its applications emphasizes the urgency of an effective approach, with reference to both the side of fear and that of hope, of ambition. With all the widespread complexity and doubt, it is becoming more and more manifest that only a concrete and integral Privacy by Design approach can provide a solution. But due to the unending series of privacy breaches, anxiety and distrust will continue to dominate.

This is a constant existential factor, as privacy is rooted in the individual's solitude. Rationally it culminates in the methodical doubt of Descartes, philosopher of the Enlightenment: distrust characterizes our attitude to life. Actually, Big Data gain for everyone presupposes a mathematical argumentation, but there is no such thing and if there was, only a minority would understand it.

The fundamental trade-off character of privacy prevents a consensus. But this is exactly where the fair play of economic potential starts for everyone. As has been noted at the beginning of this port, digital privacy is the capacity to negotiate social relationships by controlling access to personal information.

Laws, policies and technology increasingly structure people's relationships with social institutions, authorities and one another. This offers new challenges but also opportunities with regard to privacy. For this reason, a new conceptual framework needs to be created for the analysis of privacy policies on the one hand and the design and development of data processing systems on the other.

The reasoning in this context is as follows: Big Data is a reality; it is extremely valuable, but at the same time nourishes unease with relation to privacy. A proper balance needs to be created between organizations and individuals. Privacy by Design, Privacy-Enhancing Technologies, standardized legislation in addition to the corresponding responsible behavior constitute the integral approach that should enable Big Data gain for everyone.

The Introduction to this part outlines how the personal information economy works. A variety of organizations are engaged in collecting data about us all that can end up virtually anywhere, through different types of information brokers: with banks, marketers, media, government authorities, legal organizations, individuals, legal instances and employers. Only organizations that exclusively collect privacy-neutral information of fewer than 5,000 individuals a year and do not share it with third parties in any way whatsoever fall outside the scope of this ecosystem, according to the American Federal Trade Commission. All other parties need to pay serious attention to the implementation of Privacy by Design and to simple options for consumers, and they must continue to demonstrate transparency towards the market.

As privacy, data protection and personal information represent such a high economic and relational value, organizations need to operationalize Privacy by Design on the basis of the following seven basic principles:

1. **Privacy by Design means that you take proactive and preventive action: not reactive – no repairs afterwards**

Try to anticipate so-called *privacy-invasive* events as much as possible and, first and foremost, try to prevent them. Do not wait until a privacy invasion presents itself.

2. **Privacy guarantee needs to be the default setting**

You aim to guarantee maximum privacy for individuals and make sure that personal information is safe and secure in any IT system and business operation. There should be no need for individuals to worry about this or to take action.

3. **Privacy needs to be embedded in the design**

Privacy requirements need to be an integral part of the design and the architecture of IT systems and business operations. Privacy is an essential component of the functionality that is supplied.

4. **Go for full functionality: no poor trade-off but a clearly positive balance**

Address the legitimate privacy interests and objectives as a win-win situation. Avoid apparent opposites such as privacy versus security and demonstrate that they may well occur simultaneously.

5. **Solutions need to be totally conclusive and unequivocal: end-to-end security at all times**

Security is a central element. One of the aspects of data protection is that all data can be destroyed securely at the end of a process or other lifecycle, or at any desired moment.

6. **Ensure full visibility and transparency: openness is your leitmotiv**

It should be perfectly clear to stakeholders what exactly is going on with regard to all business operations and IT solutions. It should be possible for any party involved to check this at any time.

7. Deal with privacy respectfully: particularly by focusing on the individual
Strong privacy defaults, a timely explanation of what is going on, and user-friendly options for individuals are indispensable to a relationship based on mutual trust. The interaction is decisive in this context.

These principles bear upon the core of any organization: digital technology, design and infrastructure plus the operation itself. They have been further elucidated and elaborated in the report entitled *Operationalizing Privacy by Design: A Guide to Implementing Strong Privacy Practices* of December 2012, complete with actions and responsibilities in the organization on the part of management, software architects, developers, business-line owners and owners of applications. It also includes specific examples, such as the healthcare and energy sectors and, in the field of technology, camera surveillance and near-field communication (tap & go).

Privacy: great – but what's the next step...?

1. For an overall picture of digital privacy for your business, see for example the Checklist of Responsible Information-Handling Practices from the US Privacy Rights Clearinghouse. Another good starting point is the Checklist of Basic Questions about Privacy and Confidentiality from the Privacy Tool Kit (sub IV) of the American Library Association. Also, the Data Protection and Privacy Download Pages on the website of the European Committee for Standardisation CEN are rich sources of information.
2. Chapter 4 of this part suggests that you make a *Privacy Impact Assessment* (PIA).
3. The structural development of managing privacy as an economic catalyst is called *Privacy by Design* (PbD). There is a high degree of consensus concerning the benefit of this possible solution. Privacy by Design has been operationalized in the conclusion of this part on the basis of seven recommendations.

No More Secrets
Management Summary

In 2005, the term Big Data was coined by O'Reilly Media, which had presented Web 2.0 a year earlier. "Big Data" is digital data too large, complex and dynamic for conventional data tools and systems to capture, store, manage and analyze. Big Data has become an increasingly topical subject with key relevance for all other Business & Customer Technology fields from analytics, mobile and social to cloud. In terms of technology development and business adoption, the Big Data field has undergone extremely rapid changes, and that is an understatement.

In 1890, the first Big Data challenge arose. U.S. Government Census was carried out and all 60 million people needed to be counted manually. Herman Hollerith solved this challenge with his Pantograph punched-card tabulating machine. This method cut time down from ten years to les than 24 months. Another milestone before the advent of the Internet was in 1965, when the first data center was built. The U.S. government needed a place to store 742 tax returns and 175 million sets of fingerprints. All records were transferred to magnetic tape and stored on one big computer. Although the plan was aborted because of security concerns, this was the birth of the data center concept.

In **Creating Clarity with Big Data**, Part I of this book, we respond to the question of what Big Data actually is, where it differs from existing data processing practices, and how the transformative potential of Big Data can be estimated.

The desire of charting **Your Big Data Potential** arises from the purposeful data focus on the combination of business, organization and technology. In Part II of this book we list ten questions to help you formulate a concrete plan. These questions, just like the checklist at the end of this Part, have been validated by discussions with various experts and clients.

Question 1 Why Big Data intelligence?
Question 2 What new insights can I expect?
Question 3 How will these insights help me?
Question 4 What skills do I need?
Question 5 How do Big Data pioneers organize data management and IT processes?
Question 6 How can I merge my structured and unstructured data?
Question 7 Which new technologies should I be watching?
Question 8 What is looming on the horizon?
Question 9 What does this mean in organizational terms?
Question 10 How does this affect everyday life?

The concrete adoption of new policy and plans in organizations form the main theme of this book's Part III, entitled **Big Social: Predicting Behavior with Big Data.** Many concrete Big Data related adoption and plans of organizations are currently oriented toward the theme of Big Social: basically the customer side, particularly inspired by the social network activity of Web 2.0.

The data explosion is occurring all around us, but an important part of the discussion concerns the extent to which organizations ought to commit themselves to Big Data. The answer is: they should do so on the basis of well-considered policy. Policy, whether it comes from external or internal sources, has to cope with the privacy issue, which is the subject of the final Part of this book called **Privacy, Technology and the Law.**

In the introduction of this book we shed some light on **The Future of Big Data** related to the five major discussion topics we encountered in the many conversations we had. These are: acceleration, transformation, data ownership, privacy and Edward Snowden.

With **No More Secrets with Big Data Analytics** VINT aims to create clarity by putting experience and vision in perspective: independent and supported by examples. It is impossible to provide exhaustive answers to many of the issues that arise, and probably they raise even more questions – about strategic choices that you have to make.

We want to extend our discussion with you about this new Big Data frontier: online at http://vint.sogeti.com and, of course, in person. Actively participating will contribute to clear and responsible decisions via progressive insight.

Literature and Illustrations

Part I

Anderson, C. (2008): “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Appro Supercomputer Solutions (2012): “From Sensors to Supercomputers (Part 1)”, http://www.appro.com/blog/from-sensors-to-supercomputers-part-1/

Appro Supercomputer Solutions (2012): “From Sensors to Supercomputers (Part 2)”, http://www.appro.com/blog/from-sensors-to-supercomputers-part-2/

Credit Suisse Equity Research (2012): *The Apps Revolution Manifesto. Volume 1: The Technologies*, https://doc.research-and-analytics.csfb.com/docView?sourceid=em&document_id=x442413&serialid=RLlLcqX7GjvqRVYI/qyvW0PUffebK64M3spKCnjhF74%3D

Economist Intelligence Unit/SAS (2011): *Big Data: Harnessing a Game-changing Asset*, http://www.sas.com/resources/asset/SAS_BigData_final.pdf

Frost & Sullivan (2011): *Big Science > Big Data > Big Collaboration – Cancer Research in a Virtual Frontier*, http://www.emc.com/collateral/analyst-reports/fs-big-science-big-data-big-collaboration.pdf

Gartner (2012): “Information Management Goes ‘Extreme’: The Biggest Challenges for 21st Century CIOs”, http://www.sas.com/offices/NA/canada/lp/Big-Data/Extreme-Information-Management.pdf

Harbor Research (2012): “Smart Systems Drive New Innovation Modes”, http://harborresearch.com/smart-systems-drive-new-innovation-modes/

Hortonworks (2012): “7 Key Drivers for the Big Data Market”, http://hortonworks.com/blog/7-key-drivers-for-the-big-data-market/

IBM (2011): *Big Data Success Stories*, http://www-03.ibm.com/press/us/en/attachment/35525.wss?fileId=ATTACH_FILE1&fileName=IBM%20Big%20Data%20Success%20Stories.pdf

IBM Data Management (2012): “Big Data Governance: A Framework to Assess Maturity”, http://ibmdatamag.com/2012/04/big-data-governance-a-framework-to-assess-maturity/

IDC/SAS (2011): *Big Data analytics: Future architectures, Skills and roadmaps for the CIO*, http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf

Leadership Council for Information Advantage/EMC (2011): *Big Data: Big Opportunities to Create Business Value*, http://www.emc.com/microsites/cio/articles/big-data-big-opportunities/LCIA-BigData-Opportunities-Value.pdf

McKinsey Global Institute (2011): *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Mehta, C. (2012): “4 Big Data Myths – Part II”, http://cloudcomputing.blogspot.nl/2012/04/4-big-data-myths-part-ii.html

MIT Sloan Management Review/IBM Institute for Business Value (2010): *Analytics: The New Path to Value*, http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-embedding-analytics.html

Sumser, J. (2012): "Big Data: The Questions Matter Most", http://www.hrexaminer.com/big-data-the-questions-matter-most/

The 451 Group (2010): "Total data: 'bigger' than big data", http://blogs.the451group.com/information_management/2010/12/06/total-data-bigger-than-big-data/

UN Secretary-General (2011): Global Pulse, http://www.unglobalpulse.org/

Wolfram, S. (2011): "Jeopardy, IBM, and Wolfram|Alpha", http://blog.stephenwolfram.com/2011/01/jeopardy-ibm-and-wolframalpha/

World Economic Forum (2012): *Big Data, Big Impact: New Possibilities for International Development*, http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf

Yared, P. (2012): "Big Data may be hot, but 'little data' is what makes it useful", http://news.cnet.com/8301-1001_3-57424600-92/big-data-may-be-hot-but-little-data-is-what-makes-it-useful/

Part II

Capgemini (2012): "18 Reasons Why Your Organization Needs a Chief Data Officer (CDO)," http://www.capgemini.com/blog/capping-it-off/2012/06/18-reasons-why-your-organisation-needs-a-chief-data-officer-cdo

Capgemini (2013): *The Role of the Chief Data Officer in Financial Services*, http://www.capgemini.com/sites/default/files/resource/pdf/the_role_of_the_cdo_in_financial_services.pdf

Capgemini Consulting & MIT Center for Digital Business (2012): *The Digital Advantage: How Digital Leaders Outperform Their Peers in Every Industry*, http://ebooks.capgemini-consulting.com/The-Digital-Advantage

Center for Large Scale Data Systems (2012): *Defining a Taxonomy of Enterprise Data Growth*, http://storage-brain.com/wp-content/uploads/papers/Enterprise-Data-Growth-Ver-1.0-FINAL.pdf

Cisco (2013): *Big Data: Big Potential, Big Priority*, http://newsroom.cisco.com/uk/press-release-content?type=webcontent&articleId=1158061

Cisco IBSG (2012): *Unlocking Value in the Fragmented World of Big Data Analytics: How Information Infomediaries Will Create a New Data Ecosystem*, http://www.cisco.com/web/about/ac79/docs/sp/Information-Infomediaries.pdf

Cooper, M. & P. Mell (2012): "Tackling Big Data," http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf

Daley, R. (2012): "How to Be Successful with Big Data Integration," http://slashdot.org/topic/bi/how-to-be-successful-with-big-data-integration

Fisk, P. (2006): *Marketing Genius*

Franks, B. (2012): *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*

Fraunhofer IAIS (2013): *Innovationspotentialanalyse 2013*, http://www.iais.fraunhofer.de/bigdata-studie.html

IDC & EMC (2012): *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

IDC & SAS (2011): *Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO*, http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf

Inmon, W.H. & K. Krishnan (2011): *Building the Unstructured Data Warehouse*

Kalakota, R. (2012): "Organizing for BI, Analytics and Big Data: CoE, Federated or Departmental," http://practicalanalytics.wordpress.com/2012/06/19/organizing-for-bi-analytics-and-big-data/

Krishnan, K. (2012): "Big Data & Analytics," http://www.widama.us/Documents/Big-Data-June-2012.pdf

Krishnan, K. (2013): *Data Warehousing in the Age of Big Data*

Mayer-Schönberger, V. & K. Cukier (2013): *Big Data: A Revolution that Will Transform How We Live, Work, and Think*

Mayer-Schönberger, V. & K. Cukier (2013): "The Rise of Big Data," http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data

McKinsey Center for Business Technology (2012): *Perspectives on Digital Business*

McKinsey Global Institute (2011): *Big data: The next frontier for innovation, competition and productivity*, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

MIT Sloan Management Review (2013): "How Starbucks Has Gone Digital," http://sloanreview.mit.edu/article/how-starbucks-has-gone-digital

Moore, G.E. (1965): "Cramming more components onto integrated circuits," http://download.intel.com/museum/Moores_Law/Articles-Press_releases/Gordon_Moore_1965_Article.pdf

Oracle (2012): *Oracle Information Architecture: An Architect's Guide to Big Data*, http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

Palmisano, S. (2010): "Welcome to the Decade of Smart," http://www.ibm.com/smarterplanet/us/en/events/sustainable_development/12jan2010/

SAS & CMO Council (2013): *Big Data's Biggest Role: Aligning the CMO & CIO – Greater Partnership Drives Enterprise-Wide Customer Centricity*, http://www.cmocouncil.org/download-center.php?id=259

Sattel, G. (2012): "Business Models and the Singularity," http://www.digitaltonto.com/2012/business-models-and-the-singularity

Shirky, C., quoted by K. Kelly (2010): http://www.kk.org/thetechnium/archives/2010/04/the_shirky_prin.php

TCS (2013): *The Emerging Big Returns on Big Data*, http://www.mitcio.com/sites/default/files/sponsorwp/TCS-Big-Data-Global-Trend-Study-2013.pdf

University of Texas (2011): "Measuring the Business Impacts of Effective Data," http://www.wipro.com/images/infographic.jpg

Wal-Mart Labs, http://www.walmartlabs.com/platform

Warden, P. (2011): *Big Data Glossary*

Part III

AMP Lab: UC Berkeley Algorithms, Machines and People Lab, http://amplab.cs.berkeley.edu/

Anderson, C. (2008): "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Barrett, P. (2012): "10 questions CMOs need to ask about social media", http://www.asterdata.com/blog/2012/04/26/10-questions-cmos-need-to-ask-about-social-media

Cellan-Jones. R. (2012): "Facebook advertising: Who likes my virtual bagels?", http://www.bbc.co.uk/news/technology-18812126

Center for Economics and Business Research (2012): *Data equity: Unlocking the value of big data*, http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf

Chief Customer Officer Council (2012): "The Role of the CCO", http://www.ccocouncil.org/site/the-role-of-the-cco.aspx

Conan-Doyle, A. (1892): "The Adventure of the Copper Beeches", http://sherlockholmes.wikia.com/wiki/Story_Text:_The_Adventure_of_the_Copper_Beeches

Dachis, J. (2012): "Big Data Is The Future Of Marketing", http://www.businessinsider.com/big-data-is-the-future-of-marketing-2012-7

Domenico, M. De, A. Lima and M. Musolesi (2012): *Interdependence and Predictability of Human Mobility and Social Interactions*, 2012. (Read also M. C. González, C. A. Hidalgo, Barabási: *Understanding Individual Human Mobility Patterns*.)

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

Economist, The, Insurance data (2012): "Very personal finance. Marketing information offers insurers another way to analyse risk", http://www.economist.com/node/21556263

Economist, The, Special Report: International Banking (2012): "Crunching the numbers. Banks know a lot about their customers. That information may be valuable in more ways than one", http://www.economist.com/node/21554743

Etlinger, S. (2011): "Research Report: A Framework for Social Analytics", http://susanetlinger.wordpress.com/2011/08/10/research-report-a-framework-for-social-analytics, http://www.altimetergroup.com/research/reports/a-framework-for-social-analytics

Eysenbach, G. (2006): "Gunther Eysenbach coins the term 'Infodemiology' and wins AMIA award", http://www.ehealthinnovation.org/?q=node/89

Fader, P. (2012): *Customer Centricity: Focus on the Right Customers for Competitive Advantage*, http://executiveeducation.wharton.upenn.edu/resources/upload/Wharton-ExecEd-Customer-Centricity-excerpt.pdf (excerpt)

FuturICT: http://www.futurict.eu

Gartner (2010): "Gartner Identifies the Top 10 Strategic Technologies for 2011", http://www.gartner.com/it/page.jsp?id=1454221

Gartner (2010): "Next-Generation Analytics: Adding the Social Dimension", http://www.scribd.com/doc/81893261/February-9-Top-10-Strategic-Tech-Dcearley

Global Pulse: http://www.unglobalpulse.org/

Gomes, L. (2012): "Is There Big Money in Big Data?", http://www.technologyreview.com/news/427786/is-there-big-money-in-big-data/

GraphLab: GraphChi, http://graphlab.org/graphchi/

Guazelli, A. (2012), "Predicting the future ... in four parts": [1] What is Predictive Analytics? [2] Predictive modeling techniques [3] Create a predictive solution [4] Put a predictive solution to work, http://www.predictive-analytics.info/2012/07/predicting-future-in-four-parts.html

Hardy, Q. (2012): "Big Data for the Poor", http://bits.blogs.nytimes.com/2012/07/05/big-data-for-the-poor

Hausmann, V. et al. (2012): "Developing a Framework for Web Analytics", http://uni-koblenz-landau.academia.edu/PetraSchubert/Papers/1862203/Developing_a_Framework_for_Web_Analytics

Hickins, M. (2012): "Taking Small Steps to Big Data", http://blogs.wsj.com/cio/2012/05/24/taking-small-steps-to-big-data/

History of Social Media from 550 BC to 2010: http://www.webanalyticsworld.net/wp-content/uploads/blogger/10-socialMediaTL_05.png

Hubbard, D. (2007): “How to Measure Anything: Finding the Value of ‘Intangibles’ in Business, http://www.hubbardresearch.com/wp-content/uploads/2011/08/TAC-How-To-Measure-Anything.pdf,http://howtomeasureanything.com/

Ideya (2012): *Social Media Monitoring Tools and Services Report 2012*, http://ideya.eu.com/images/SMMTools ReportExcerpts 09072012Final.pdf

McIntyre, S. (2012): “From Reactive to Predictive Analytics”, http://www.radian6.com/blog/2011/08/from-reactive-to-predictive-analytics/

McKinsey Global Institute (2011): *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

Mejova, Y. (2009): ‘Sentiment Analysis: An Overview’, http://homepage.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf

MyBuys (2012): “MyBuys Named the Leader in Personalization for Third Year in a Row”, http://finance.yahoo.com/news/mybuys-named-leader-personalization-third-110000414.html

Nerny, C. (2012): “Point Smartphone, Get Data: IBM to Unveil Augmented Reality App for Retailers”, http://data-informed.com/point-smartphone-get-data-ibm-to-unveil-augmented-reality-app-for-retail

Nokia Mobile Data Challenge, http://research.nokia.com/page/12000

Police 2.0: The Possibilities of the Digital Revolution for the Dutch Police Force (in Dutch), http://criminaliteitswijzer.ning.com/

Recorded Future: Unlock The Predictive Power Of The Web, https://www.recordedfuture.com/

Rezab, J. (2012): “70% of Fans Are Being Ignored By Companies – Now what?”, http://www.socialbakers.com/blog/655-70-of-fans-are-being-ignored-by-companies-now-what

Roebuck, K. (2011): *Social Analytics: High-impact Emerging Technology – What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*

SAS (2012): “Claims Fraud: Prevent fraud before claims are paid”, http://www.sas.com/industry/ins/fraud.html

SAS, UN (2012), “Can a country’s online ‘mood’ predict unemployment spikes?”, http://www.sas.com/news/preleases/un-sma.html

Shaw, W. (2012): “Cash machine: Could Wonga transform personal finance?”, http://www.wired.co.uk/magazine/archive/2011/06/features/wonga

Talbot, D. (2012): “A Phone That Knows Where You’re Going”, http://www.technologyreview.com/news/428441/a-phone-that-knows-where-youre-going/

Ungerleider, N. (2011): “The Federal Reserve Plans To Monitor Facebook, Twitter, Google News”, http://www.fastcompany.com/1786730/federal-reserve-plans-monitor-facebook-twitter-google-news

VINT (2007): *Me the Media: Rise of the Conversation Society*, http://www.methemedia.com/

VINT (2012): “Creating clarity with Big Data”, http://blog.vint.sogeti.com/?p=4559

Wallace, M. (2011): “Social Analytics is more than just Social Media…”, http://allthingsanalytics.com/2011/11/04/social-analytics-is-more-than-just-social

Wallace, M. (2011): “What’s in a name?”, http://allthingsanalytics.com/2011/10/13/whats-in-a-name

Walmart Labs: “Social Genome”, http://www.walmartlabs.com/social/social-genome/

Webtrends/DK New Media (2011): “History of Web and Social Analytics”, http://www.marketingtechblog.com/infographic-history-web-social-analytics/

Wiki: Social Media Listening, Monitoring, Measuring, and Management Tools, http://socialmedia-listening.wikispaces.com/Tools

Wiki: A Wiki of Social Media Monitoring Solutions, http://wiki.kenburbary.com/social-meda-monitoring-wiki

Part IV

Agre, P.E. & M. Rotenberg (1997): *Technology and Privacy: The New Landscape*, http://polaris.gseis.ucla.edu/pagre/landscape.html

Alessandro Acquisti, A. (2010): “The Economics of Personal Data and The Economics of Privacy,” http://www.oecd.org/sti/interneteconomy/46968784.pdf, http://www.heinz.cmu.edu/~acquisti/papers/acquisti_privacy_economics.ppt

Alvaro, A. (2012): *Lifecycle Data Protection Management: A contribution on how to adjust European data protection to the needs of the 21st century*, http://www.alexander-alvaro.de/wp-content/uploads/2012/10/Alexander-Alvaro-LIFECYCLE-DATA-PROTECTION-MANAGEMENT.pdf

American Library Association (ca 2005): Privacy Tool Kit / Checklist of Basic Questions about Privacy and Confidentiality, http://www.ala.org/offices/oif/iftoolkits/toolkitsprivacy/guidelinesfordevelopingalibraryprivacypolicy/guidelinesprivacypolicy

Article 29 Data Protection Working Party (2013): “European data protection authorities publish their joint opinion on mobile apps,” http://ec.europa.eu/justice/data-protection/article-29/press-material/press-release/art29_press_material/20130314_pr_apps_mobile_en.pdf

Asimov, I. (1951): *Foundations*

Bradbury, R. (1951): *Fahrenheit 451*

Burkert, H. (1997): “Privacy-Enhancing Technologies: Typology, Vision, Critique,” http://books.google.nl/books?id=H2KB2DK4w78C&pg=PA125

Bygrave, L.A. (2002): “Privacy-Enhancing Technologies – Caught between a Rock and a Hard Place,” http://folk.uio.no/lee/publications/PETS_speech.pdf

Cavoukian, A. (2009): “Privacy by Design: The 7 Foundational Principles,” http://www.privacybydesign.ca/content/uploads/2009/08/7foundationalprinciples.pdf

Cavoukian, A. (2012): *Operationalizing Privacy by Design. A Guide to Implementing Strong Privacy Practices*, http://privacybydesign.ca/content/uploads/2012/12/operationalizing-pbd-guide.pdf

Cavoukian, A. & J. Jonas (2012): *Privacy by Design in the Age of Big Data*, http://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf

CEN: European Committee for Standardisation (2010): Data Protection and Privacy Download Pages, http://www.cen.eu/cen/Sectors/Sectors/ISSS/CWAdownload/Pages/DPPCWA.aspx

Center for Internet and Society PET wiki: http://cyberlaw.stanford.edu/wiki/index.php/PET

Clarke, R. (1995-2013): Dataveillance & Information Privacy, http://www.rogerclarke.com/DV

Clarke, R. (2001): “Introducing PITs and PETs: Technologies Affecting Privacy,” http://www.rogerclarke.com/DV/PITsPETs.html

Clinton, W.J. & A. Gore (1997): “A Framework voor Global Electronic Commerce,” http://clinton4.nara.gov/WH/New/Commerce/read.html

Cloud Security Alliance (2012): *Top Ten Big Data Security and Privacy Challenges*, https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf

Computers, Privacy & Data Protection (2013): “Reloading Data Protection,” http://www.cpdpconferences.org/sponsors.html

Coueignoux, P. (2006): “Liabilities and Vulnerabilities in the Information Age,” http://www.eprivacy.com/lectures/toc.html

Coueignoux, P. (2012): “The Privacy Tax. Open letter to Barack Obama and Herman van Rompuy,” http://www.cawa.fr/the-privacy-tax-article005879.html

Coueignoux, P. (2013): ePrio, trustworthy social eCommerce, http://eprivacy.com

Cyberspace Law and Policy Centre (2008): *Distinguishing PETs from PITs: Developing technology with privacy in mind*, http://www.cyberlawcentre.org/ipp/publications/papers/ALRC_DP72_Technology_final.pdf

Daniel P. (2013): “Abine’s DeleteMe app review: deal with personal info databases from the comfort of your phone,” http://www.phonearena.com/news/Abines-DeleteMe-app-review-deal-with-personal-info-databases-from-the-comfort-of-your-phone_id38722

De Wereld Draait Door (2012): “Doorstart EPD: Wilna Wind en Alexander Klöpping,” http://dewerelddraaitdoor.vara.nl/media/197890

Department of Defense (2013): Personally Identifiable Information Course Module, http://iase.disa.mil/eta/pii/pii_module/pii_module/index.html

Department of Health, Education and Welfare (1973): “The Code of Fair Information Practices,” http://epic.org/privacy/consumer/code_fair_info.html

Department of Homeland Security (2011): *Handbook for Safeguarding Sensitive Personally Identifiable Information At The DHS*, http://www.dhs.gov/xlibrary/assets/privacy/privacy_guide_spii_handbook.pdf

Diploma of Information Technology, Knowledge Management (2012): "Types of Privacy," http://toolboxes.flexiblelearning.net.au/demosites/series4/411/content/privacy/types_of_privacy.htm

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

Edelman (2012): "Privacy & Security: The New Drivers of Brand, Reputation and Action. Global Insights 2012," http://edelmaneditions.com/wp-content/uploads/2012/03/Data-Security-Privacy-Infographic_Final.png

Electronics Weekly (2010): "Electroon #9 - 'Computer says No'" (1960), http://www.electronicsweekly.com/blogs/electronics-weekly-blog/2010/09/electroon-9---computer-says-no.html

Ellis Smith, R. (2004): *Ben Franklin's Web Site: Privacy and Curiosity from Plymouth Rock to the Internet*, http://www.privacyjournal.net/_center_ben_franklin_s_web_site__privacy_and_curiosity_from_plymouth_rock_to_the_3087.htm

ENISA (2012): *Study on monetising privacy: An economic model for pricing personal information*, http://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/monetising-privacy

European Commission (2010,2011): *Special Eurobarometer 359: Attitudes on Data Protection and Electronic Identity in the European Union*, http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf

Europese Commissie (2012): *Voorstel voor een Verordening van het Europees Parlement en de Raad betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens (algemene verordening gegevensbescherming)*, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:NL:PDF http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf

Export.gov (2012): U.S.-EU & U.S.-Swiss Safe Harbor Frameworks, http://export.gov/safeharbor

Fair Information Practice Principles, FIPS (1973, 1980), http://simson.net/ref/2004/csg357/handouts/01_fips.pdf

Federal Trade Commission (2012): Fair Information Practice Principles, http://www.ftc.gov/reports/privacy3/fairinfo.shtm

Federal Trade Commission (2012): "Marketing Your Mobile App: Get It Right from the Start," http://business.ftc.gov/documents/bus81-marketing-your-mobile-app

Federal Trade Commission (2012): *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policy Makers*, http://www.ftc.gov/os/2012/03/120326privacyreport.pdf

Fielder, A. (2013): "European Parliament committees threaten wholesale destruction of privacy and data protection rights," https://www.privacyinternational.org/blog/european-parliament-committees-threaten-wholesale-destruction-of-privacy-and-data-protection

Foremski, T. (2011): "SFCurators: Our Public, Private, And Secret Lives…," http://www.siliconvalleywatcher.com/mt/archives/2011/07/sfcurators_our.php

Frankfurter Allgemeine (2011): "Der deutsche Staatstrojaner wurde geknackt," http://www.faz.net/aktuell/chaos-computer-club-der-deutsche-staatstrojaner-wurde-geknackt-11486538.html

Galexia (2008): "The US Safe Harbor – Fact or Fiction?," http://www.galexia.com/public/research/assets/safe_harbor_fact_or_fiction_2008/safe_harbor_fact_or_fiction-Recommen.html

Garfinkel, S. (2001): *Database Nation. The Death of Privacy in the 21*st *Century*

Gerber, B. (2009, 2010): OECD Privacy Principles, http://oecdprivacy.org

Gorodyansky, D. (2013): "It's Data Privacy Day: 3 Things You Must Do," http://www.inc.com/david-gorodyansky/3-ways-to-go-all-out-this-data-privacy-day.html

Greenberg, A. (2008): "The Privacy Paradox," http://www.forbes.com/2008/02/15/search-privacy-ask-tech-security-cx_ag_0215search.html

Hagel, J. (2012): "The Rise of Vendor Relationship Management," http://edgeperspectives.typepad.com/edge_perspectives/2012/06/the-rise-of-vendor-relationship-management.html

Hamlin, K. (2012): "Personal Data List in Mind Map Form," http://www.identitywoman.net/personal-data-list-in-mind-map-form

Hirschleifer, J. (1979): *Privacy: Its Origin, Function, and Future*, http://www.econ.ucla.edu/workingpapers/wp166.pdf

HP Laboratories (2011): *Privacy-Enhancing Technologies: A Review*, http://www.hpl.hp.com/techreports/2011/HPL-2011-113.pdf

Huffington Post (2010): "Facebook's Zuckerberg Says Privacy No Longer A 'Social Norm'" (video), http://www.huffingtonpost.com/2010/01/11/facebooks-zuckerberg-the_n_417969.html

Human Rights Council (2012): Resolution A/HRC/20/L.13: The promotion, protection and enjoyment of human rights on the Internet, http://daccess-dds-ny.un.org/doc/UNDOC/LTD/G12/147/10/PDF/G1214710.pdf

Huxley, A. (1932): *Brave New World*

Information Commissioner's Office (1998): "Data protection principles," http://www.ico.gov.uk/for_organisations/data_protection/the_guide/the_principles.aspx

Information Commissioner's Office (2013): "Privacy impact assessment (PIA)," http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/privacy_impact_assessment.aspx

Internationaal privacykader (2013): http://www.cbpweb.nl/Pages/ind_wetten_int.aspx

IT Law Wiki (2013): "Privacy-Enhancing Technologies" [incl. UK ICO & EU], http://itlaw.wikia.com/wiki/Privacy%E2%80%90enhancing_technologies

IT Law Wiki (2013): "The IT Law Wiki," http://itlaw.wikia.com/wiki/The_IT_Law_Wiki

Johnson, H. (2013): "The Application Privacy, Protection, and Security (APPS) Act of 2013," http://apprights-hankjohnson.house.gov/2013/01/apps-act.shtml

Kennard, W.E. (2012): "Remarks by U.S. Ambassador to the EU, William E. Kennard, at Forum Europe's 3rd Annual European Data Protection and Privacy Conference," http://useu.usmission.gov/kennard_120412.html

Koorn, R.F. & J. ter Hart (2011): "Privacy by Design: From privacy policy to privacy-enhancing technologies," http://www.compact.nl/artikelen/C-2011-0-Koorn.htm

Kozinski, A. (2012): "The Dead Past," http://www.stanfordlawreview.org/online/privacy-paradox/dead-past

Kuner, C. et al. (2012): "The Challenge of Big Data for Data Protection," http://idpl.oxfordjournals.org/content/2/2/47.extract

Kuneva, M. (2009): Keynote Speech, Roundtable on Online Data Collection, Targeting and Profiling, http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm

Lee, F. (2006): *An Investigation of Privacy Tradeoff on the Internet*, http://citebm.business.illinois.edu/twc%20class/project_reports_spring2006/privacy%20issues/lee/internet_privacy_fei.pdf

Lynley, N. (2013): "A Palantir Founder Suggests His Startup Is Worth About $8 Billion," http://blogs.wsj.com/digits/2013/01/16/a-palantir-founder-suggests-his-startup-is-worth-about-8-billion

Mackay, S. (2012): "'Apps' and Big Data and Privacy – An Oxymoron?," http://ediscoverytalk.blogs.xerox.com/2012/12/10/apps-and-big-data-and-privacy-an-oxymoron

Magenta Advisory (2012): "Wise use of consumer data enables to improve companies' performance and creates possibilities for new services and solutions," http://www.magentaadvisory.com/2012/09/12/wise-use-of-consumer-data-enables-to-improve-companies-performance-and-creates-possibilities-for-new-services-and-solutions/

Microsoft (2012): *Differential Privacy for Everyone*, http://www.microsoft.com/en-us/download/details.aspx?id=35409

Microsoft HealthVault: http://www.healthvault.me/, https://www.healthvault.com/nl/nl, http://www.microsoft.com/health/en-us/products/Pages/healthvault.aspx

Microsoft HealthVault Ecosystem: http://www.netsoft-usa.com/images/img_medtracker_HealthVaultFuture.png

MozillaWiki (2011): Privacy Icons project (beta release), https://wiki.mozilla.org/Privacy_Icons

Mulligan, D. (2012): "Bridging the Gap between Privacy and Design," http://www.law.berkeley.edu/14542.htm

Nader, R. (1965): "Unsafe at Any Speed," http://www.nndb.com/people/788/000023719/

National Institute of Standards and Technology (2010): *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

OECD (2013): "Information security and privacy," http://www.oecd.org/sti/interneteconomy/informationsecurityandprivacy.htm

Olsson, M. (2012): "NSA Building A $2 Billion Quantum Computer Artificial Intelligence Spy Center," http://mind-computer.com/2012/05/13/nsa-building-a-2-billion-quantum-computer-artificial-intelligence-spy-center

Öman, S. (2004): *Implementing Data Protection in Law*, http://www.scandinavianlaw.se/pdf/47-18.pdf

OpenLearn (2012): "Secret or sharing? Play our Privacy Game," http://www.open.edu/openlearn/privacy

Orwell, G. (1948): *1984*

Out-law.com (2012): "Smart meter technology is privacy intrusive," http://www.out-law.com/en/articles/2012/january-/smart-meter-technology-is-privacy-intrusive-researchers-claim

Packard, V. (1964): *The Naked Society*, http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=3842&context=fss_papers

Pfanner. E. (2013): "French Tax Proposal Zeroes In on Web Giants' Data Harvest," http://www.nytimes.com/2013/02/25/technology/french-tax-proposal-zeroes-in-on-web-giants-data-harvest.html

PISA Consortium (2003): *Handbook of Privacy and Privacy-Enhancing Technologies: The case of Intelligent Software Agents*, http://www.cbpweb.nl/downloads_technologie/pisa_handboek.pdf

Pratt, W.F. (1979): *Privacy in Britain*, http://books.google.nl/books?id=GDjNgEgw2fgC

"Privacy & Free Speech: It's Good for Business," http://www.dotrights.org/business/primer

Privacy by Design (2013), http://www.privacybydesign.ca

Privacy Impact Assessment .nl (2013): Privacy Quick Scan, http://privacyimpactassessment.nl/QuickScan.html

Privacy Rights Clearinghouse (2013): "Empowering Customers. Protecting Privacy." Fact Sheet 12: Checklist of Responsible Information-Handling Practices, https://www.privacyrights.org/fs/fs12-infohandling.htm, https://www.privacyrights.org/about_us.htm

Privacy, Technology and the Law, http://www.judiciary.senate.gov/about/subcommittees/privacytechnology.cfm

Rand, A. (1943): *The Fountainhead*

Reyburn, S. (2012): "ACT debuts the App Privacy Icons," http://www.insidemobileapps.com/2012/10/04/act-debuts-the-app-privacy-icons/

Rosen, J. (2012): "The Right to Be Forgotten," http://www.stanfordlawreview.org/online/privacy-paradox/right-to-be-forgotten

Rousseau, J-J. (1754): *Discourse on the Origin and Basis of Inequality among Men*

RT, Russia Today (2012): "NSA refuses to disclose its links with Google," http://rt.com/usa/nsa-epic-foia-court-413

RTL Z (2013): "Privacy op internet niet goed beschermd," http://www.rtl.nl/components/financien/rtlz/nieuws/2013/03/privacy-op-internet-niet-goed-beschermd.xml

Rubinstein, I. (2011): *Regulating Privacy By Design*, https://www.privacyassociation.org/media/pdf/knowledge_center/Regulating_privacy_by_design.pdf

Rubinstein, I. (2012, 2013): *Big Data: The End of Privacy or a New Beginning?*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2157659

Schoen, S. (2009): "What Information is 'Personally Identifiable'?," https://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable

Searls, D. (2010): "Do we have to 'trade off' privacy?," http://blogs.law.harvard.edu/vrm/2010/09/19/do-we-have-to-trade-off-privacy

Sengupta, S. (2012): "Building an Iconography for Digital Privacy," http://bits.blogs.nytimes.com/2012/11/19/building-an-iconography-for-digital-privacy

Smith (2012): "Digital privacy in the big data era: Microsoft's data protection keynote," http://www.networkworld.com/community/blog/digital-privacy-big-data-era-microsofts-data-protection-keynote

Sogeti Executive Summit 2013, held in October in Amsterdam, http://vint.sogeti.com/tag/exsum13/, http://automotive.tomtom.com/events/webinars/automobile.html, http://www.tomtom.com/en_gb/congestionindex/, https://www.tomtom.com/lv_lv/legal/privacy/, http://www.tomtom.com/en_gb/safeguarding-your-data/

Sogeti VINT (2012, 2013): vier Big Data-onderzoeksnotities, http://vint.sogeti.com/bigdata

Solove, D.J. (2006): *A Taxonomy of Privacy*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=667622

Solove, D.J. (2011): *Nothing to Hide: the False Trade-off between Privacy and Security*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1827982

Solove, D.J. & P.M. Schwartz. (2011): *Privacy Law Fundamentals*, https://www.privacyassociation.org/media/pdf/publications/PLF_TOC.pdf, "Chapter 1: An Overview of Privacy Law in all its varied types and forms and a timeline with key points in the development of privacy law," https://www.privacyassociation.org/media/pdf/publications/PLF_Chap_1.pdf

South Australian Law Reform Institute (2012): *Computer says no. Modernisation of South Australian evidence law to deal with new technologies*, http://www.law.adelaide.edu.au/reform/downloads/issues-paper-1-computer-says-no.pdf

Spiegel, Der (2012): “Surfing for Details: German Agency to Mine Facebook to Assess Creditworthiness,” http://www.spiegel.de/international/germany/german-credit-agency-plans-to-analyze-individual-facebook-pages-a-837539.html

Strahilevitz, L.J. (2004): *A Social Networks Theory of Privacy*, http://www.law.uchicago.edu/files/files/230-ljs-privacy.pdf

Surveillance Studies Network (2006): *A Report on the Surveillance Society for the Information Commissioner*, http://www.surveillance-studies.net/?page_id=3

Szoka, B. (2009): *Privacy Trade-Offs: How Further Regulation Could Diminish Consumer Choice, Raise Prices, Quash Digital Innovation & Curtail Free Speech*, http://ftc.gov/os/comments/privacyroundtable/544506-00035.pdf

Tavani, H. & D. Vance (1996): “Chapter 4.3 Computers and Privacy,” http://home.aisnet.org/displaycommon.cfm?an=1&subarticlenbr=633

Tene, O. & J. Polonetsky (2012): *Big Data for All: Privacy and User Control in the Age of Analytics*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2149364

Tene, O. & J. Polonetsky (2012): “Privacy in the Age of Big Data: A Time for Big Decisions,” http://www.stanfordlawreview.org/online/privacy-paradox/big-data

TNO, TILT (2011): *Trusted Technology. Een onderzoek naar de toepassingsvoorwaarden voor Privacy by Design in de elektronische dienstverlening van de overheid*, http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/12/05/trusted-technology-een-onderzoek-naar-de-toepassingsvoorwaarden-voor-privacy-by-design-in-de-elektronische-dienstverlening-van-de-overheid.html

Tompor, S. (1994): “The Credit Report from Hell,” http://www.recordnet.com/apps/pbcs.dll/article?AID=/19940801/A_NEWS/308019321

TrendLabs (2012): *Be Privy to Online Privacy*, http://about-threats.trendmicro.com/ebooks/be-privy-to-online-privacy/files/assets/downloads/publication.pdf

Upshure, R.E.G. et al. (2001): “The privacy paradox: laying Orwell’s ghost to rest,” http://www.ncbi.nlm.nih.gov/pmc/articles/PMC81333

Verenigde Naties (1948): Universele Verklaring van de Rechten van de Mens, artikel 12, http://www.un.org/en/documents/udhr/index.shtml#a12

Vitaliev, D. (2011): “Data Protection and Privacy,” http://dmitri.vitaliev.info/data-protection-and-privacy

VNO NCW (2011): Brief: Kabinetsnotitie Privacy met o.m. Privacy Ouick Scan (PIA), http://www.vno-ncw.nl/SiteCollectionDocuments/Brieven/brief11-11507.pdf

Warren, S. & L. Brandeis (1890): “The Right to Privacy,” http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html

Westin, A.F. (1967): “Privacy and Freedom,” http://scholarlycommons.law.wlu.edu/cgi/viewcontent.cgi?article=3659&context=wlulr

Westin, A.F. & M.A. Baker (1972): *Databanks in a Free Society. Computers, Record-keeping, and Privacy*

Wet bescherming persoonsgegevens (2013), http://www.cbpweb.nl/pages/ind_wetten_wbp.aspx

White House, The (2012): *Consumer Data Privacy in a Networked World. A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

Wolfensberger, D.R. (2006): "Congress and the Right to Privacy," http://www.wilsoncenter.org/sites/default/files/privacy-essay-drw4.pdf

World Economic Forum (2011): *Personal Data: The Emergence of a New Asset Class*, http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf

World Economic Forum, Boston Consulting Group (2013): *Unlocking the Value of Personal Data: From Collection to Usage*, http://www.weforum.org/issues/rethinking-personal-data

Zamyatin, Y. (1924): *We*

VINT.SOGETI.COM

About Sogeti

Sogeti is a leading provider of professional technology services, specializing in Application Management, Infrastructure Management, High-Tech Engineering and Testing. Working closely with its clients, Sogeti enables them to leverage technological innovation and achieve maximum results. Sogeti brings together more than 20,000 professionals in 15 countries and is present in over 100 locations in Europe, the US and India. For more information please visit www.sogeti.com.

About VINT

It is an arduous undertaking to attempt to keep up on all developments in the IT field. The state-of-the-art IT opportunities are often very far from the workings of the core business. Sources that provide a deeper understanding, a pragmatic approach, and potential uses for these developments are few and far between. VINT, a Sogeti research institute, provides a meaningful interpretation of the connection between business processes and new developments in IT.

A balance is struck between factual description and intended utilization in every report drawn up by VINT for the investigations it carries out. VINT uses this approach to inspire organizations to consider and use new technology.

VINT | Vision • Inspiration • Navigation • Trends

ISBN 978-90-75414-73-8
< 9789075 414738 >