GSR discussion paper

Big Data - Opportunity or Threat?

Work in progress, for discussion purposes

Comments are welcome!

Please send your comments on this paper at: gsr@itu.int by 20 June 2014.

The views expressed in this paper are those of the authors and do not necessarily reflect the opinions of ITU or its Membership.

Table of Contents

Page

Big Data - Opportunity or Threat 1

Executive summary 1

1. The opening 2

2. Setting the Stage 2

What is and what drives Big Data? 3

How much and how fast is the data in the world growing? 5

Big Data's defining qualities 6

Eight Principles 8

3. The opportunities 10

For the enterprise 10

For the individual; as a consumer; as a citizen 12

In medicine. 12

For society 13

4. Outcomes with Concern 14

A New Dark Side 16

Erosion of Trust & Big Data Backlash 17

False confidence in Data 18

The Rise of the Data Barons 18

5. The Role of (and future for) Regulatory Authorities 19

Ensure Protection 19

Protecting Human Volition / Taming Probabilistic Predictions 19

Facilitating Publicly Available Big Data Expertise 20

Keeping Data Markets Fluid 20

6. Forums, discussions and papers 21

BIG – Big Data Public Private Forum 21

The World Economic Forum 21

The International Telecommunication Union 22

US: White House, Big Data Initiative 24

7. The Wrap 24

Annex 1 – 15 Worst Data Breaches (2000-2012) 25

©ITU 2014

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Big Data - Opportunity or Threat

Authors: Mr. Andrew J Haire & Dr. Viktor Mayer-Schönberger

Executive summary

Big Data offers a new perspective on reality, and therefore will affect and shape all sectors of our economy, especially those that play a role in the capturing and/or relaying of data and information.
But Big Data's likely impact is broader than the economy; it affects how our societies make sense of the world, and decide important policy challenges, and as you will read, innovation.

This paper is divided into four parts: initially it provides some boundaries to the subject; next, the contributions that Big Data offers to society and individuals are explained; as a balance, the attention of reader is drawn to some of the inherent risks of this powerful new technological tool; and finally, it concludes with the regulatory and policy considerations that should be accounted for when crafting future policy.

We draw the reader's attention to the conclusion as an area for focus for establishing policy and the rules that will encourage the further use and benefits derived from Big Data, but to set the proper frameworks to prevent abuses, be they societal or individual.

1. The opening

Google can predict the spread of the seasonal flu from Internet search queries it receives. Airplane engine manufacturers can predict when an engine part will break before it actually does, allowing that part to be changed at a convenient time and place rather than when the airplane is in mid-flight. A startup company, *Inrix* offers a smartphone app that helps about one hundred million users every working day to drive from home to work and back, avoiding heavy traffic in real time. And a Dutch mobile phone operator discovered that changes in the signal strength of cell towers could be translated into local weather data, thus giving the operator a potentially lucrative and very comprehensive network of thousands of weather stations capturing real-time data.

All these are examples of Big Data; our ability to gain insights from large amounts of data that would not be attainable from much smaller amounts, and that in turn leads not only to higher efficiency but to innovative new products and services. Much like in other instances an increase in quantity results in a change in quality. We have seen this in the past, too. If one takes a photo of a horse galloping across the field every minute, then they are still just photos. But if one takes a photo every sixteenth of a second, and shows the resulting images in fast succession, the vast increase in the quantity of captured information translates into a new quality: film; and an industry was born. Something similar is happening with Big Data.

Big Data in essence offers a new perspective on reality, and therefore will affect and shape all sectors of our economy, especially those that play a role in the capturing and/or relaying of data and information.

But Big Data's likely impact is broader than the economy; it affects how our societies make sense of the world, and decide important policy challenges, and as you will read, innovate.

2. Setting the Stage

The world of Big Data over the past few years has rapidly evolved both in the marketplace and in the research community. As in many other areas, the rules governing Big Data has been slow to adapt. Further, *what* we think we knew just a few years back is now either changed or refined. The intent of this paper is to offer a foundation, showing what Big Data is, explaining where we've been, and looking at where ICT regulators, policy makers and other authorities, such as Competition Authorities or Data Protection Authorities, let's collectively call them *Regulatory Authorities*, have set or should set some boundaries. We hope to provide an understanding to foster a stronger appreciation both nationally and globally of the makeup of this term Big Data and the points of light that make up the discussion.

This paper is divided into four parts: initially we will give some boundaries to the subject; next, we explain the contributions that Big Data offers to society and individuals; as a balance, we would like to draw attention to some of the inherent risks of this powerful new technological tool; and finally, we will conclude with the regulatory and policy considerations that should be accounted for when crafting future policy.

More specifically, this is a paper about Big Data, and its characteristics, its history, its future and most importantly what Regulatory Authorities – as defined earlier - can and should do to meet its challenges without dampening opportunities. As regulators, what can, or should, be done to carry out your mandate of responsibilities? We remain sensitive that no two countries or economies share a common or identical governance structure to oversee the tech or media or other societal sectors, so we will treat them having a similar mandate, to keep the paper's discussion understandable. Reflexive actions by policy makers often lead to individual's protections rights, possibly at the expense of individual's opportunities. We hope to offer this discussion that will allow the reader to find that balance, taking into account the needs and character of their particular jurisdiction. We will further treat, for the purpose of this paper, the Regulatory Authority as one holding a responsibility to promote market health, growth, and opportunity, but with a role to protect those who rightfully can't protect themselves.

We will try to present enough diversity in the practicalities and uses of Big Data to offer awareness between the benefits and the risks; but place it in a context that allows for understanding where the industry has been and where it could be going.

Big Data obviously is closely connected to our ability to gather, analyze, and store data easily and relatively at a low cost. Therefore, most accept that there are two fundamental drivers why Big Data has arrived. The cost of computing (both processing and storage) has dropped, and the ease at which we communicate has risen. Add to this the vast amount of research in both academic and corporate communities to better connect what seems to be 'unrelated data' to becoming 'related'.

What is and what drives Big Data?

In computing, Moore's Law described a doubling of computing power roughly every eighteen month at constant cost. That means one can get double the performance or half the price after 1.5 years. Moore's Law has been observed for over fifty years now, and while eventually Moore's Law will hit hard physical limits sometime in the 2020s for current technologies, further paradigm breaking computing technologies are being investigated that would push these limits out much further.

Progress similar to Moore's Law can be observed with storage density and storage cost. In fact, in recent years storage cost for some digital storage media has dropped even faster than computing cost. Thus, data that cost USD150,000 to store in 1970 now costs USD0.01. As a result storing digital information is now very affordable on very fast devices.[1]

Additionally, our software tools to manage digital storage have vastly improved, providing very fast retrieval times. But that is only half the story. The other half is the rise of a whole new breed of databases over the last fifteen years or so that are capable of storing very diverse and unstructured data, rather than the highly curated and finely structured data records of the 1980s. These rather messy unstructured databases, such as Google's *MapReduce* or the open-sourced *HADOOP* are now mature, providing fast storage and easy retrieval, with over ½ of the *Fortune 50*[2] using this platform.

Taken together more often than ever before in human history we now can decide by default to keep and store data that we have collected rather than to purge it soon after collection, because storage is now affordable and data retrieval keeps huge data holdings accessible. But two more phenomena are contributing to the current data deluge. The first is vastly improving sensor technology, making it possible to render ever more aspects of human existence into data format, precisely and at low cost. Only two decades ago, capturing location (for instance through GPS receivers) was a costly affair. Today the chips are so cheap, and can be augmented with other location technologies for vastly improved accuracy. And users are embracing this newfound capacity. The chart above demonstrates that three quarters of smartphone owners get directions from their phone – thereby agreeing to indicate where they are[3]. But a far fewer subset of that group use a service to find their friends.

[1] Pingdom.com. February 2010, http://royal.pingdom.com/2010/02/18/amazing-facts-and-figures-about-the-evolution-of-hard-disk-drives/

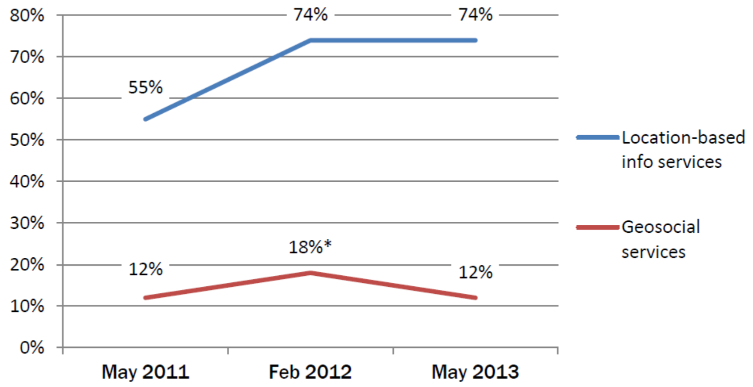
[2] PR Newswire; Altior's AltraSTAR – Hadoop Storage Accelerator…, 18 Dec 2012. Retrieved 1 May 14.

[3] http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_Location-based%20services%202013.pdf

Use of location-based information and geosocial services among smartphone owners, 2011-2013

For location services: % of smartphone owners who use their phone to get directions, recommendations, or other information related to a location where they happen to be.

For geosocial services: % of smartphone owners who use a service such as Foursquare or Gowalla to "check in" to certain locations or share their location with friends.



* Slight wording change since previous survey

Source: Pew Research Center's Internet & American Life Project tracking surveys. For 2011 data, n=2,277 adults ages 18 and older. For 2012, n=2,253 adults. For 2013, n=2,252 adults. All surveys were conducted via landline and cell phone, in English and Spanish.

Sensors are now also available for everything from movement and acceleration to environmental aspects (temperature, atmospheric pressure, UV exposure), to the now booming field of health (heart rate, blood oxygenation, even blood sugar levels). To demonstrate versatility, in the mid-2000s sensors were placed on experimental basketballs to calculate spin, location and trajectory – and more importantly, would the shot 'go in', and if not, why not. Soon sensors will go even further, capturing aspects such as smell with far greater precision than today. Other sensors capture vibration, weight, distention and many other aspects of physical properties.

The most versatile sensors, so to speak, of course are humans themselves. Revealing data about them or even more importantly about others on social networking sites, through fitness, health and quantified-self data platforms account for another substantial portion of increased data streams available.

While much of the data we create we believe is evident and even viewable (we write an email and we see the results), we also leave behind our transparent fingerprints everywhere we go. We have a phone that knows our movements, but what about the surveillance camera that image-identifies us; the airport scanner that knows what we travel with; the credit card that knows our eating habits down to the food we like? Our cars know where we drive, when we drive, and how fast; our library card knows what we read and view; our health monitor (if we choose to own one) knows where and when we walk, run, cycle – and even what our heart rate is. And all of this data can be saved, stored, communicated and under some circumstances, shared. In sum, this data where you've been, where you are, and now where you might be going, is being collected at rates faster than ever before.

In a novel experiment, and subsequent research paper[4] three members of the Computer Science Department at the University of Rochester, (New York, USA) explored the relationship between people's location, their interactions and their social ties through a social network. Amazingly, even if you deliberately created your online-self as "dark" – making yourself private and invisible – this analytical approach (referred in their work as 'Flap') would be able to predict your physical location within 100 meters with an accuracy of 47 percent.

As mentioned above, the second phenomenon that contributes is networking speed and reach. Data bandwidth is continuously increasing by leaps and bounds throughout the world, both in wired and

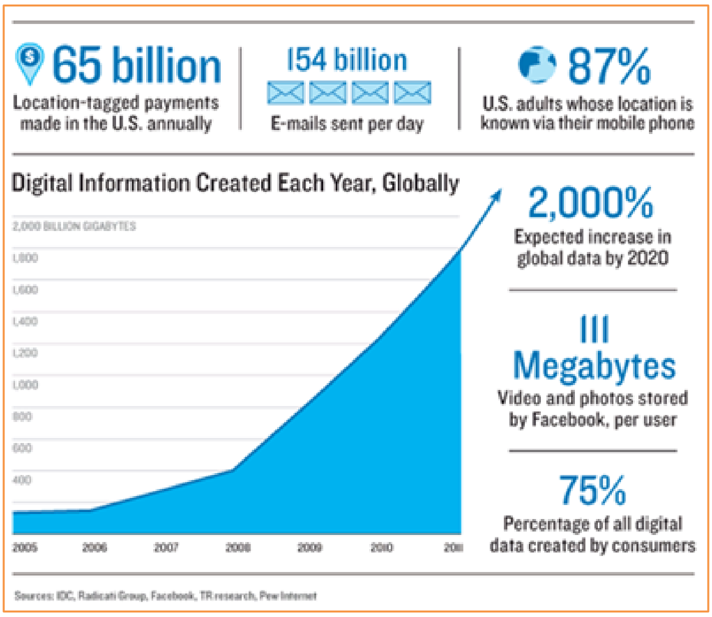
[4] Adam Sadilek, Henry Kautz and Jeffrey Bingham, "Finding Your Friends and Following Them to Where You Are", 5th ACM Conference on Web Search and Data Mining, 2012

wireless networks. LTE wireless networks rolled out in many large metropolitan areas around the globe provide what used to be broadband speeds available to wired networks only a few years ago. Moreover, relatively recent backbone and undersea cable activity has connected geographic areas around the world to the Internet that have long been underserved. East Africa is a particularly salient case in point here. As networks become more powerful, and reach further, more and more data can be shared, exchanged, but more importantly combined and analyzed together to further advance Big Data insights.

How much and how fast is the data in the world growing?

The best "guestimates" of the total amount of data in the world suggest that from 1987 to 2007 the total amount of analog and digital data in the world grew from 3 billion gigabytes to 300 billion gigabytes, a 100x increase in two decades.[5] In research by Cisco, the computer manufacturer has added color to the profound scope of the data that exists and is being created.[6]

Many of these statistics have found their way into discussions over the past few years, but none is more telling that ninety percent of the world's data has been created in the past two years.

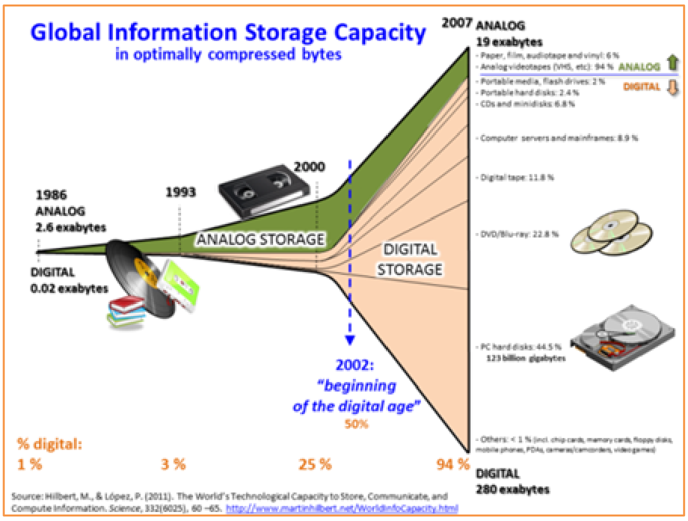


Sources: IDC, Radicati Group, Facebook, TR research, Pew Internet

[5] Martin Hilbert and Priscilla López, "The world's Technological Capacity to Store, Communicate, and Compute information." *Science* 1 (April 2011), pp. 60–65

[6] Cisco. Cisco Visual Networking Index; http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

There is another shift taking place. Data that was in the past stored in an analogue format and not necessarily ready for analytics, is now in digital form, and this creates huge opportunities for analysis, indexing, mining. Voice, video and other visual content can be more efficiently diagnosed and analyzed and indexed for mining and identification with the other digital indices and of kept data. Historically this analogue medium grew slowly, held a relatively short shelf life, aged quickly, and provided an infrequent means to connect with existing digital data. The common and familiar storage formats were tape cassettes, vinyl records, and celluloid film. Sound was (and to a degree still is) analogue. Today, sound and images – Skype and YouTube, to name just two - are digitized before they are transmitted. But this all is changing. In the year 2000, three quarters of data had been analog, now more than 99 percent of data in the world is digital.



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Data growth has been accelerating recently. More than 90 percent of all data that exists was created in the last two years.[7] IDC, a research firm, predicts that there will be 44 times more digital data by the end of 2020 that there was in 2009 - or put differently - the amount of digital data doubles every 20 months. It is not just people that are creating data. A Boeing 777 airplane generates a terabyte of data during a three hour flight; and after 20 such flights it has generated more data than presently is in the world's largest library, and as technology improves the aircraft will be capable of capturing up to 30 terabytes from its sensors[8]. Today 75% of data is created by individuals with activities such as emails, documents, downloading movies, to name a few.

Throughout the world, in a growing number of governments projects are under way to make vast troves of data collected by government publicly available so that individuals but also companies can use it. Often termed "open data" these initiatives not only aim to improve public as well as accountability democratic deliberation and participation through increased transparency. Governments also see "open data" as a non-monetary way to incentivize and facilitate big data entrepreneurship and innovation. It is a "data subsidy" instead of the more traditional (and much more costly) monetary subsidy and has led to literally thousands of applications around the globe.

Additionally, the World Wide Web Foundation (www.webfoundation.org), based in Switzerland is fostering engagements with stakeholders to further develop Open Government Data (OGD) initiatives in low and middle income countries. The intent of these initiatives are to improve transparency and accountability, and by do so increase the efficiency and effectiveness of government.

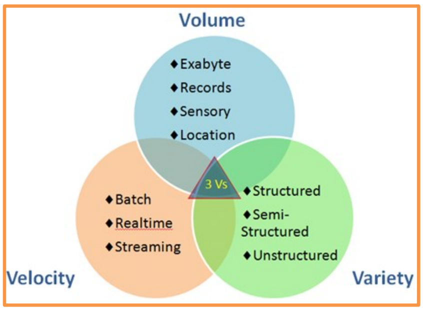
Particularly Internet companies are drowning in data. Over one hundred million photos for instance are uploaded to Facebook every single hour, much like an hour of video on YouTube every second. And Google is said to process well over a petabyte of data every single day – that is the entire amount of all data stored in the largest library of the world, the US Library of Congress, one hundred times over.

Big Data's defining qualities

[7] SINTEF. "Big Data, for better or worse: 90% of world's data generated over last two years." ScienceDaily. ScienceDaily, 22 May 2013. <www.sciencedaily.com/releases/2013/05/130522085217.htm>.

[8] Rosenbush, Steve. The Wall Street Journal, CIO Journal. 13 November 2013.

So it is tempting to look at this deluge of data that is accelerating and think of it, and its drivers as capturing and being the essence of Big Data. Research firms and other corporate stakeholders, such as Gartner, McKinsey and IBM, and institutions such as the ITU have put forwards a plentitude of acronyms and labels to encapsulate the Big Data qualities, such as the often used three Vs, of (high) volume, (high) velocity, and (high) variety. There is a fourth V offered: veracity – or simply the believability of the data itself. We would also like to draw attention to the ITU's Technology Watch Report (November) 2013 that further address the meaning and uses of Big Data. But we believe that the definition here lends itself to a more qualified definition, if for no other reason than the Big Data landscape has evolved so dramatically recently, so this will be explored further below.



But we suggest that these terms fail to capture what Big Data is really all about. To understand Big Data, we need to understand how humans have made sense of the world so far. For millennia we have done so by observing the world, and gaining insights from our observations. For hundreds of years, we have systematically captured data, and evaluated it to reveal ever more details about reality. But capturing data always had been extraordinarily costly and difficult, and so was analyzing and storing data.

So to save cost and time, humans have devised methods and mechanisms, institutions and processes to answer questions by collecting and analyzing as little data as was absolutely necessary to do so. Because of cost, we chose to live and think in a world of Small Data – and we understand reality based on this constraint.

If the constraint of capturing, analyzing and storing data goes away, we can rethink our deeply rooted traditions of how we make sense of the world around us. This is what Big Data is all about: it is a new lens on reality, capturing not just a simplified version of it that gave us a first (but often blurry) glimpse, but a detailed version that captures and illuminates reality comprehensively and in its full complexity.

Hence, the defining qualities of Big Data are deeper and more profound than what often is suggested. The three terms that capture this are: *more, messy and correlations*.

- **More**: This means that we can now capture and analyse more data *relative* to the phenomenon we want to understand, the question we want to answer than before when we relied on small samples and subsets of data. That way we can look into details with unprecedented clarity, and even more importantly we can answer questions that we did not even think of when we collected the data (which is often impossible when just relying on a data sample). This is what experts mean when the say that we can now "let the data speak". So what counts is not the absolute number of data points (the "volume"), but the relative number of data points that captures and let's see reality as it is.
- **Messy**: In the times of Small Data, we spent a lot of effort ensuring that the limited number of data points we cared to capture and analyse were of high quality. That is understandable. If you only have 100 data points, getting 20 of them wrong will skew the result, leading to bad consequences, what is sometimes called GIGO – "garbage in, garbage out". But in the age of Big Data our ability to capture many magnitudes more of data, will make it more cost-effective to go for more data even if the data is of varying quality than to expend great cost at capturing little data at high quality. It is not that we give up on exactitude entirely; it is only that we give up our singular devotion to it. What we possibly lose on the micro level, we gain in insight at the macro level.
- **Correlations**: Humans always thrive to find causes for what they observe and experience. This comforts us and gives us the sense that we understand the world. But often the causes we identify are simply wrong. Statisticians have long made the point that with most statistical analysis we are not able to tease out causalities, but correlations – seeming connections within

the data. Correlations do not tell us why things are happening, but they tell us what is happening, and that already can be an important insight.

For instance, large retailer Wal-Mart through a Big Data analysis of transaction data discovered that before a hurricane, people buy batteries and flashlights, as one would expect. But they also discovered through correlational analysis that people bought Pop Tarts, a sugary snack. For Wal-Mart it does not matter why people buy Pop Tarts – but it is very valuable to know that people are buying Pop tarts before a storm. That way, Pop Tarts can be moved to a more prominent location in the store, and more of them are sold. Similarly Amazon does not know why certain people buy certain products together with others, but know that they buy such products drives Amazon's product recommendation engine, and is said to be responsible for about 35 percent of Amazon's revenues.[9]

This does not make the quest for causal linkages superfluous, but it strongly implies that rather than venturing into often incorrect assumptions and suggestions of "why", we are better advised to use correlational analysis to first understand *what* is going on. That in it may sometimes be good enough, full of valuable insights that drives innovation. But it also acts as a powerful filter to highlight what specific correlations we may want to investigate further to understand the underlying causes, making such explorations far more cost-effective than ever before.

Taken together, more and messy data, analyzed often first through identifying correlations gives us a very unique, very powerful, and comprehensive lens into reality, and thus let's make predictions about the present and the future. In short, this is what at its core Big Data is all about.

Derived from these defining qualities of Big Data, and in line with the drivers at play that enable Big Data as outlined above, the core economic principle of Big Data comes into focus. It is not that data can provide insights – humans have known that for millennia. It is that as we move from an age of Small Data to an age of Big Data, what we do with data and how we extract value, especially economic value from it changes.

In the Small Data age, not only was relatively little data collected, but it was gathered with a specific purpose in mind. Once the purpose was fulfilled, the data had achieved its value, and often was put aside, forgotten, or at times even actively purged because of high storage cost.

In the age of Big Data the value of data is not exhausted by applying the data to the purpose for which the data was collected. Rather the value of data is the sum of the many uses and reuses the data can be put to that might not have been obvious at the time of collection, but turn out to reveal insights that are worth a lot.

Eight Principles

This has huge consequences on how commercial entities collect, analyze and store data, which can be summarized in these eight general principles:

- **Data Retention**: In the Big Data age it makes sense to store data much after it has fulfilled its original purpose, because it might still hold value that can be extracted by reusing it for novel purposes. For instance, *Google* looks at old search queries to learn what mistakes people make when typing words, and thus is able to correct these mistakes automatically, leading to what arguably is the world's best spell checker.
- **Data Collection**: In the Big Data age, it may make sense for commercial entities positioned at the flow of data to capture and store that data even if it cannot be used for a particular purpose yet as the data may hold dormant value. As an example, *Facebook* is saving years and years of user input because that data holds latent value even though currently *Facebook* dos not fully extract that value.

[9] Matt Marshall, Aggregate Knowledge raises $5m from Kleiner, on a roll; VB News, December 10, 2006.

- **Data Primacy**: It is data that holds value, and so those that have data or have access to it will be able to extract that value, while those that do not have data will suffer economically. This is the real meaning of the shorthand that data is the "new gold" or "new oil". That analogy is actually insufficient, as unlike physical resources such as gold or oil, the potent value claim with data is that its value is not exhausted by being used once. Unlike physical resources it can be recycled many, many times over and still provide value.

- **Data Expertise**: The expertise to extract the hidden value in data is very important for commercial entities, and currently there is a shortage of experts in this field. These data scientists are therefore in high demand and are able to command high salaries. Eventually however this will change as the labour markets adjust to this demand by creating an increasing supply of data scientists, much as they did in the past with telecom experts, software programmers, network engineers, or web designers.

- **Data Mindset**: More important arguably than the technical expertise in analyzing big data sets is the strategic ability to see value in specific data, and to be focused on exploiting that. This is one of the reasons why a small (but growing) cadre of Big Data entrepreneurs has had serial successes in Big Data start-ups entering even relatively crowded market spaces. Professor Oren Etzioni, who founded travel price forecaster *Farecast* and consumer goods forecaster *decide.com* is an excellent example.

- **Non-linear Scalability of Data**: Because data's value increases with the possible connections between data points, having more data will disproportionally increase data's value. It is Big Data's network effect, and it means that scale efficiencies are neither linear nor step-linear, but following a power law. More data means much more value. This will drive many large Big Data companies to become even larger data holders.

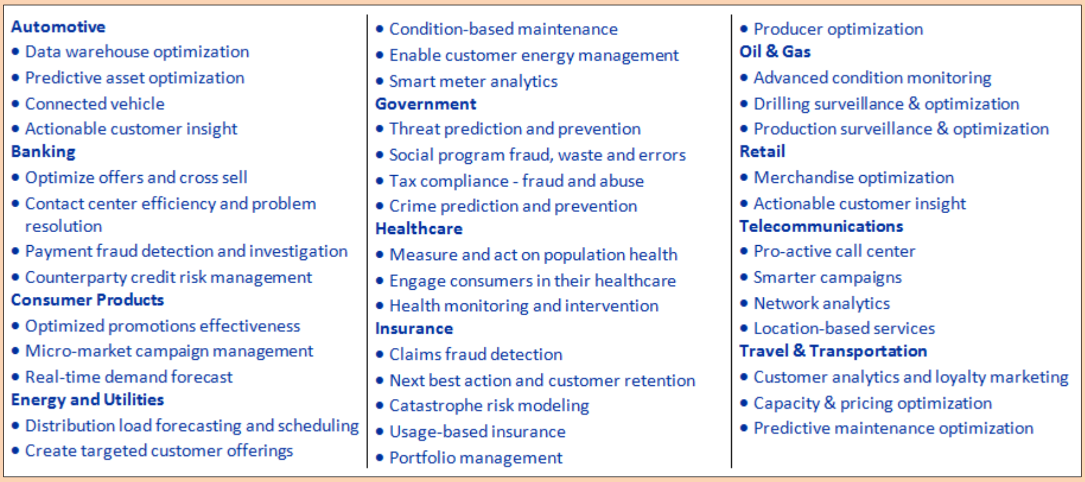
- **Reduced Barriers to Entry**: At the same token, a Big Data start-up does not have to invest heavily in technical infrastructure to process and store data. Unlike in the previous generation of start-ups, such as Facebook and Google, Big Data start-ups can utilize cloud storage and cloud processing capacity that provides them with flexible commodity priced capabilities when they need it. This greatly reduces the barriers to entry, and creates strong incentives for companies and entrepreneurs to begin utilizing Big Data. So while the big may become bigger, the small and nimble retain a very strong proposition to succeed.

- **Data's Utility**: The utility of data will be irrespective of the economic sector the data's holder is operating in. So for instance, a telecom operator might find it to be a weather data platform, or a car manufacturer may turn itself into a data platform for mobility and travel. This means that companies with traditional revenue streams and in established sectors may both find themselves capable through their ability to capture and analyse data to enter other sectors and add new revenue streams, as well as also find themselves competing against new entrants or those from completely other sectors.

Improving efficiency is of course very important for any economic player, and particularly important for players in sectors that offer largely commoditized products, such as in telecommunications. Lowering the cost of production of a product or service is essential, as businesses in these sectors struggle to stay profitable.

It is obvious and understandable given the powerful nature of Big Data that the initial focus of Big Data on businesses and business models has been its impact on efficiency. For instance, a quick glance of applications that a major computer company, IBM, is tackling emphasizes efficiency:

Figure 1: Enterprise applications with a focus on Big Data



Automotive
- Data warehouse optimization
- Predictive asset optimization
- Connected vehicle
- Actionable customer insight

Banking
- Optimize offers and cross sell
- Contact center efficiency and problem resolution
- Payment fraud detection and investigation
- Counterparty credit risk management

Consumer Products
- Optimized promotions effectiveness
- Micro-market campaign management
- Real-time demand forecast

Energy and Utilities
- Distribution load forecasting and scheduling
- Create targeted customer offerings
- Condition-based maintenance
- Enable customer energy management
- Smart meter analytics

Government
- Threat prediction and prevention
- Social program fraud, waste and errors
- Tax compliance - fraud and abuse
- Crime prediction and prevention

Healthcare
- Measure and act on population health
- Engage consumers in their healthcare
- Health monitoring and intervention

Insurance
- Claims fraud detection
- Next best action and customer retention
- Catastrophe risk modeling
- Usage-based insurance
- Portfolio management
- Producer optimization

Oil & Gas
- Advanced condition monitoring
- Drilling surveillance & optimization
- Production surveillance & optimization

Retail
- Merchandise optimization
- Actionable customer insight

Telecommunications
- Pro-active call center
- Smarter campaigns
- Network analytics
- Location-based services

Travel & Transportation
- Customer analytics and loyalty marketing
- Capacity & pricing optimization
- Predictive maintenance optimization

Source: IBM

An efficiency strategy alone, however, is not going to be a long-term solution, as efficiencies only give relative advantages vis-à-vis the competition, but are not creating new revenue streams (and thus business value) themselves. Fortunately, the real power and role of Big Data is not limited to enhancing efficiencies. In fact, it goes far beyond that. Big Data creates new insights that will enable new products and services. Big Data is, perhaps more than anything else, a tool for innovation. The role of data thus changes, from an auxiliary function of enabling efficient transactions to becoming itself valuable, and thus turning into a source of revenue and profit. This we will explore further in the following section.

3. The opportunities

Big Data offers a great number of opportunities, which we canvass in this section. Depending on who is benefitting primarily from these opportunities, we have this section divided in thirds: a section for opportunities in the enterprise, for opportunities for the individual and opportunities for society at large.

For the enterprise

McKinsey & Company, a consultancy, reported[10] that “Big Data” generates significant financial value across: US health care; EU public sector administration; Personal location data; Retail productivity; Manufacturing. This report spells out that enterprises need to prepare for what is coming; not just go along for a ride. Their findings followed with: data has swept into the industry landscape and it has become as important as capital and labor; ways exist to have data create value (transparency; more drives accuracy; more gives greater segmentation; improved decision making; product development); it will become the competitive edge – more in some sectors than others; use will provide productivity and efficiency; there will be a shortage of talent.

Further research by MGI and McKinsey had outlined where to uncover or unlock data within an organization that could transform into value. The interactive website[11] effectively shows where expertise has been ‘invested’ based on industry and role. The key point remains that the skill to leverage and then interpret this data is in short supply.

[10] McKinsey Global Institute; http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation p8.

[11] http://www.mckinsey.com/tools/Wrappers/Wrapper.aspx?sid={EA4CDB7F-4500-49DE-B5FB-CD35AA13806D}&pid={A30FD727-0384-4CF1-9364-4C17E9ADB9F8}

Human Resource functions. In another example, and based on an article in the SHRM's publication[12] Human Resource responsibilities from training to integrating a newly acquired company can see potential from the analytics of Big Data. HR activity has always been data-driven but with the dawn of vast amounts of additional data arriving from social media, analytics, smartphones the HR role is challenged with using this data effectively and at the same time respecting the personal and private nature – and in some cases contextual nature - of its content. Many firms now use analytics in prescreening applicants for new positions, and the simple wording chosen for a CV makes the difference between 'filed for the future' and the next round of interviews.

Telecommunications companies have not only vast amounts of operational and customer data but hold a reach insofar as their networks are local, regional and global. One firm discovered that a variation in radio frequency transmissions at mobile base stations - data already received - preceded weather changes, and became a good predictor of pending weather. Much has been made of mobile providers sharing high concentrations of their users at, say a public sporting event, to alert transport and public safety authorities of pending congestion, but in some cases anticipating where that congestion happen (which road, what form of public transport). With Big Data we can predict with a strong degree of accuracy the who will be using public transport, but where they will be going – thus allowing for a smooth degree of capacity planning. In another area of extending the usefulness of networks, undersea cable systems are built with extensive monitoring equipment to detect seismic activity. Given that scientists believe seismic waves are the most powerful tool to study the earth structure, and that 2/3 of the earth's surface is covered by ocean, these systems make a strong tool to complement study and hopefully improve prediction of future earthquake activity.

Changes to business models.

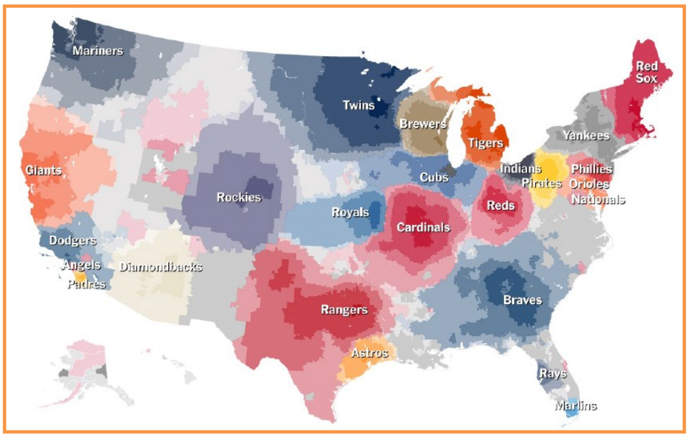
Advertising, as we once knew it, is dead. The approach of repeatedly showing images and pictures to consumers to reinforce or sell a brand is gone – probably forever. In its place are methods are highly targeted messages, friends recommendations, search analysis that result in a profoundly more efficient and effect way to reach the exact consumer the company CEO's complained that there was no clear connection between ad spending and the resulting sales; often thinking that the former was highly inefficient. Many believed that better places existed to invest their precious capital. Once that 'clicks' replaced TV viewing surveys about two decades ago, the CEOs started to gain the precision they so desperately wanted. Tools like Google *AdSense* and other integrated advertising platforms followed you from site to site, as you browsed the web – they not only know what your interests, but can observe how long you remain engaged – thus showing desire.

Recently the *New York Times*[13] printed an article on a Facebook experiment to answer a vexing question that has swirled around the US for over a century: "which is your favorite baseball team", or put differently: "where are the fans". Of course the issue is not cosmic, but it showed the power and precision of Big Data.

This is a classic consumer research question, but in the past a survey might have tried to gain insights from several thousand people, but this particular study it reached many millions of followers, allowing a far more significant degree of accuracy. To the street or even the postcode, the loyalties of were now known, but what was more valuable it showed which team was the 2nd, 3rd, 4th, etc. followed. The owner of a team knows exactly where fans are, and more importantly aren't. More importantly, it knows gender, age, buying habits, and viewing habits – it also knows if these habits are changing. You need not waste your advertising budget to reach possible fans who are already are your fans. The difference between the old and the new is accuracy, primarily from the size of the sample, and the ease as which someone can identify what they like and don't like.

[12] Bill Roberts; Society for Human Resource Management; Vol 58, No 10. 1 October 2013; http://www.shrm.org/Publications/hrmagazine/EditorialContent/2013/1013/Pages/1013-big-data.aspx

[13] New York Times, Up Close on Baseball's Borders, 24 April 2014; http://www.nytimes.com/interactive/2014/04/23/upshot/24-upshot-baseball.html



The map on the right[14], the darker shading is deeper fan interest, is available with its deep precision and accuracy on the internet to anyone who is so inclined to take a look.

For the individual; as a consumer; as a citizen

More to store and save. Individuals, too, benefit from Big Data. For instance, more efficient production of goods or services will enable companies to compete more successfully in the marketplace, including on price. This will enable individuals to get the same service at lower price, or quantitatively more at the same price (or some combination thereof). One example of this is the continuously increasing size of free email inboxes with large free email providers. Google for instance now offers 15 GB of free space to individuals, but started out with just 1 GB when it began operations in 2004.

Less road traffic. One of the easiest and in some cases powerful platforms for Big Data is to have individuals become the real-time source of data; something referred to as crowd sourcing. One such company, known as *Waze* (www.waze.com), offers individuals a clear and quick path to avoid heavy traffic. After informing *Waze* of your destination it routes you based on the known speed of others on the same route using the app. The others supply continuous driving speed for the road selected for dynamic routing for you, and re-routing if need be. One of the many clever side-benefits arrives when users warn of heavy traffic, a disabled car, or even hidden police traps – all of which will be broadcast to others using the app as they approach the area.

A small example: ever wonder how your supermarket offers coupons to you on items you have not bought today, but might be of interest? If you belong to the markets loyalty club, they know your buying habits (history), they know what you bought today (checkout), they may even know their overstock (inventory), and they know your location in their store. Bring the four together, and in real time offer you that coupon as you push your cart down the aisle past what they want you to buy.

As Big Data facilitates innovation, and thus new products and services to be developed and deployed, individuals benefit from new as well as improved products and services. In short: consumers benefit from innovation fueled by Big Data.

In medicine.

We have touched on the point that the value of data increases when more data is collected. This phenomenon is becoming quite evident in the field of medical care. The challenge lies with the tension that personal medical information is often viewed as quite private and personal, but the societal value of sharing information collectively is enormous. Medical researchers are continually looking for statistical significance in their work – but are blocked by achieving consent from the individual, so the costs of each data point remain high (often exceeding USD200 for each point). This then leverages the cost to develop medication or even medical procedures in the millions and sometimes billions of US dollars.

There should be strong incentives, mostly through public awareness and direct participation, that sharing of medical history, under controlled circumstances can yield significant public gains. In a report by the World Economic Forum it was demonstrated that by engaging individuals in a trusted way that significant improvements were achieved among the population: an 18% increase in the control of diabetes; a 20%

[14] New York Times. 24 April 2014. http://www.nytimes.com/interactive/2014/04/23/upshot/24-upshot-baseball.html

increase in the control of cholesterol and a marked difference in clinical outcomes in hospital performance where its data was published and shared.

SARS. Some may recall in 2003 the unknown nature of both the SARS infection and how it spread. With this uncertainty the public in infected areas became obsessed by avoiding contact with anyone. In addition to the tragic consequences associated with SARS, it had far reaching economic devastation. Faced with the prospect of an early symptom, the public had little choice but to precisely retrace where they recently had been, and more importantly, who they were in contact with – no easy task if you walked on a crowded street. Quarantines were often put in place for no other reason that a suspected infected person may have visited or walked nearby. Workplaces were deserted; commerce came to a halt. Special contact centers were established to provide a clearing house to find others you may have come in contact with.

Big Data, as we are starting to know now could now play a very useful role in tracking the movements of infected persons, and permitting society at large to be far better informed, and hopefully less alarmed, than a decade ago.

Research. In another area where medical research can be improved is with pediatric medicine[15]. Presently in pediatric intensive care units in hospitals measurements of patients is relatively limited, both in what is recorded and how often it is recorded. Under a significant research project in Southern California in the U.S., work is being done to greatly expand the data points using sensors on the children. A major part of this project is to start 'mining' archived pediatric data with real-time data hopefully allowing doctors access to far better analytical research, leading to improved predictive medicine for patients needing rapid diagnosis.

For society

Climate Change. Today, one of the pressing global issues (and debates) is global climate change. The data collected in not only the historic facts about our earth – temperatures of air and sea, currents, - but present day observations from weather stations throughout the world. While there isn't agreement about the future of climate changes, there is almost universal agreement that it is caused by mankind – which of course means changes, even small ones, for the good might remain in man's control. The analytical models to predict change depend heavily on Big Data, and the sharing of this data. Scientists, unlike Mr. Maury, will need to rely that this data, no matter where and how collected, remains available in an unrestricted form. Advancements in science depend on this sharing.

Education. Online learning has been available for quite some time. In the last five years substantial research and application has taken this platform to a wider and broader level. Two computer science professors from Stanford (a U.S. university), Andrew Ng and Daphne Koller founded a for-profit company offering what is called, massive open online courses – or MOOCs[16]. The departure with this company is that courses are provided at no charge, the material is of a world class nature, and the number of students that have taken courses is measured in the hundreds of thousands – and each course contains video lectures, exercises, and occasional quizzes. The courses are offered over a 6-10 week period, but each course has been designed to insure the student interacts with the material, not the traditional approach - the other way around. Keystroke biometrics is used to check the identities of enrolled students – and the effectiveness of learning the material. Peers grade homework - and statistical methods are used to complete that student's assessment; so the most important byproduct – this given a unique view into human learning given and that the sample size is so large - is higher quality education.

Crime Prevention. Predictive technologies associated with Big Data are starting to play a significant role in determining an individual's propensity to commit a crime. Errors are costly, and in some cases illegal by civil authorities. The U.S. cities of Memphis and Los Angeles are 'experimenting' with technologies

[15] Phys.org, October 22, 2012, "Using Big Data to Save Lives", http://phys.org/news/2012-10-big.html

[16] TED; June 2012; http://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education

that can determine crime 'hot spots' before they become 'hot'. Richmond, Virginia, used software to analyze crime patterns, in one case the sensor reporting of random gunfire a few hours before and after midnight on New Year's Eve. The result allowed public safety officials to find and confiscate a disproportionate number of unlicensed firearms; which resulted in removing street guns, which in turn resulted in fewer gun related offences, meaning fewer police officers necessary.

Legal side-effects. The use of new technology, be it sensors – some undetectable, GPS trackers, CCTV image scans, will start to figure in litigation, especially the question of the individual's legitimate expectation over his/her privacy. This will test the boundaries of the present evidentiary process.

Altered perceptions. Authorities often face the dilemma of having to render important policy decisions with very limited data. The results are not just ill fated public sector projects and initiatives, but a general distrust in government and the public sector. Fewer people believe government has the capacity to tackle complex policy challenges.

The opportunity beckoning with Big Data is that not only empirical data can be gathered, but so much data can be gathered that we can see how society as a whole behaves in real-time, we can watch how societal dynamics unfold at scale.

For instance, the public sector in the UK working with Big Data startup *Inrix* reuses navigation and traffic data gleaned from a large number of drivers to see commuter traffic patterns around London, and to retune their planned extensions of public transport and Park&Ride facilities. Or the Centers for Disease Control in the US have worked with Google to better understand the spread of the flu in close to real time using search queries sent to Google. Or a startup company of economists building on a research project developed at MIT capture billions of prices from eCommerce companies every day to predict in close to real time changes in consumer prices, and thus inflationary effects. So good (and objective) are their results that the Economist uses their measure instead of the official inflation rate for Argentina.

Much more is possible, and if employed correctly could greatly aid and inform public sector decision-making, and thus improve government and benefit all of us in society.

4. Outcomes with Concern

Unfortunately, but perhaps unsurprisingly there are dark sides to such a dramatic increase in the collection and storage of data, including personal data. The most obvious among them is that Big Data will result in a comprehensive infrastructure of surveillance against which even Orwell's dystopia "1984" pales in comparison.

Are you for sale? If you use your handphone, social media, subscribe to almost anything, pay with a credit card, or put another way, live in the 21st century you are creating substantial personal data about yourself. Where might this data end up? While the brands of service providers such as Verizon or AT&T may be familiar to US consumers there is a very large company, Acxiom which may not be. Acxiom collects information on about ½ a billion people around the world, and have about 1,500 data points on their behavior – or put differently about 6 billion data points available to their clients.[17] While most telecom companies globally can't sell individual information without running afoul of privacy protections, they can aggregate a grouping of individual information or delete personal identifiers from records, believing that this "anonymization" has removed the personal individuality from the data.

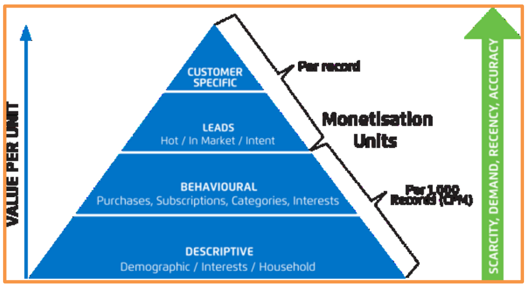
education level
estimated net worth
types of recent purchases
investments
habits (like smoking & gambling)
YOU FOR SALE
number of kids
contact info
religious & political views
marital status
mortgage amount
salary

Source : New York Times

[17] Tucker, Patrick. The Naked Future, 2014, Current. Page 119.

This practice caught the eye of American legislators about a year ago, and during a hearing before the U.S. Congress an Acxiom executive agreed to make information his company sells reviewable to the individual whose information it was. It further agreed to allow that individual to 'opt out' from having private information shared; but led the CEO to share[18] that if 20% opts out, it "would devastate our business". That alone speaks volumes about the business model of individual's surveillance. In response to these promises a website was created (AboutTheData.com) which allows a look at what they have about you.

Its value. The value of your personal information sold by marketing firms like Acxiom is directly related to how specific it can be about you. Simply they will not be able to monetize it until they collect substantial data points about you – it is far more important to know that. Herein lies the marketer's incentive – maximize the data on the individual. This creates a policy collision with the privacy rights advocates. The chart on the right demonstrates this pressure.



Not only large data brokers like Acxiom that have created a huge system of ingesting, storing and keeping ready for retrieval detailed data for hundreds of millions of people around the world, but many large global brick-and-mortar businesses, such as Wal-Mart, Target and Amazon, have done similarly for their customers. And while some of them only slowly are awakening to the commercial benefits of Big Data, once they do, they may turn into formidable powers of surveillance.

Internet companies, as one would expect, have kept personal data of their users and customers for years. Amazon is said to have captured every buying transaction, even every product a customer looked at but did not buy since its early days in the 1990s. Google is said to have kept and stored every single search query it ever received (plus data to identify from whom it came). Google receives almost a half a billion such requests every single day.[19] And Facebook is said to hold in storage well over one thousand billion data points of its one billion users; more than a half a petabyte of new data arrives at Facebook each day[20] (50 times more than the world's largest library). Parenthetically, it is reported that there is US$4 billion in potential sales abandoned each year in online shopping carts; thus creating a huge future marketing opportunity to re-mine the data – 63% of that by some estimates[21] – a value larger than the entire economy of a small country.

In addition to commercial entities, government agencies, too, are amassing huge piles of personal data, as the revelations of Edward Snowden revealed in 2013. And through data retention laws targeted at telecommunication companies, combined with sometimes opaque mechanisms, even more personal data is collected and held in storage which government agencies are able to access.

How open is a website? If you ever care to question the degree with which a website shares your information, check out PrivacyScore.com (http://privacyscore.com/). This analytic provides a score 0 to 100 showing the degree that you will be tracked and the extent that they will share your information.

Anonymization. In recent years there has been much effort to define policies to anonymize personal data. It is largely believed that these efforts will not work, mostly because we are now capturing more data, and we have stronger tools to combine and connect data. Two much publicized cases one involving Netflix, a US movie rental service, and the other, AOL[22] showed that with even basic technology someone

[18] Natasha Singer, "A Data Broker offers a Peek Behind the Curtain", New York Times, 31 August 2013.

[19] Craig Smith. DMR, 2 February 2014. By the numbers: 40 amazing Google Search Statistics and Facts.

[20] Facebook; Under the Hood, 8 November 2012; www.facebook.com/notes/facebook-engineering/under-the-hood-scheduling-mapreduce-jobs-more-efficiently-with-corona/10151142560538920

[21] Smith, Cooper. Business Insider; "Shopping Cart Abandonment…", 15May2014. http://www.businessinsider.com/heres-how-retailers-can-reduce-shopping-cart-abandonment-and-recoup-billions-of-dollars-in-lost-sales-2014-4

[22] Mayer-Schönberger, Viktor and Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think (HMH, 2013)

could re- anonymize data that these providers were convinced otherwise. Professor Paul Ohm, of the University of Colorado Law School (in the US) and expert on the harm done by de-anonymization explains in an article published in the UCLA Law Review[23] that no easy fix is available – and even arrives at the point that given enough data, no anonymization is possible because any connection makes those of us seeking anonymity an unrealistic objective. In the era of Big Data the three core strategies to insure privacy – individual notice and consent, opting out, anonymization – have lost much of their effectiveness. As mentioned earlier in this paper, researchers at the University of Rochester can identify those who chose to be 'dark' online only to 50 percent accuracy.

Given plummeting collection and storage costs, this tendency to surveil and store the captured data will likely only increase in the years to come.

A New Dark Side

But in addition to surveillance a new dark side looms, and one that so far often overlooked. This is the tempting possibility to employ big data to predict the future of an individual and hold that individual accountable for that predicted behavior. It is the idea that humans are being punished not for what they have done, but for what are only predicted to do. This may sound futuristic, and indeed is the main plot line of the 2002 Hollywood blockbuster "*Minority Report*". But it is far more science than fiction. For instance, in more than half of US states, parole boards deciding whether an incarcerated criminal should be freed on parole are utilizing Big Data analysis that portends to predict the likelihood of that criminal to be involved in a homicide in the next twelve months[24]. Dozens of police forces in US cities and metropolitan areas use "predictive policing", a Big Data analysis that forecasts when and where the next crime will be committed.

Not only government agencies employ Big Data predictions to make decisions over whom to control and to punish, commercial entities, too, use probabilistic predictions sometimes to assign individual responsibility irrespective of actual behavior. For instance, car insurers in some countries charge drivers who had bad grades in school more than those that did well in school (their prediction says that people with bad grades in schools are comparatively lousy drivers). Similarly, some people are denied loans and credit not for what they have done, but what a Big Data analysis predicts they will do (namely to default on their loan payments), even though they have never missed a payment in their past. Such behavior may be risk optimizing for the commercial entity employing it, but for the individuals affected it feels like punishment for something they have yet to do.

Society, too, is not immune to abusing Big Data in this fashion. In a world of wide-spread genetic sequencing, one can easily foresee societal pressure on individuals with genetic defects to eat or live differently, so that they minimize their chances of getting sick far into the future.

While the goals of prevention may be laudable, such a use of Big Data in effect will deny individuals human volition, their ability to act freely. If we were to accept the widespread use of probabilistic predictions for the assignment of individual responsibility we would surrender perhaps the most central individual freedom to collective fiat. It would be a different world, in which free will and individual responsibility has been marginalized (after all, if Big Data calculates who is guilty, and thus denies humans that they can decide, we cannot hold them responsible).

It is important to keep in mind, however, that the problem here is not Big Data itself, but how probabilistic predictions from Big Data analyses are being employed. Most Big Data analyses is based on correlations, on seeing seeming connections in the data that tell us "what" is happening, but do not tell us anything about the "why", about causes. In our society based on justice and freedom, individual responsibility and punishment is irrevocably linked to causality. Only those that caused others harm can

[23] Ohm, Paul. "Broken Promises of Privacy", 57 UCLA Law Review 1701 (2010)

[24] Ibid. Mayer-Schönberger, Viktor and Cukier, Kenneth p.158. Further information: Eric Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (Wiley 2013)

be held responsible, for instance. Thus it is a blatant abuse of Big Data, when one takes correlational results of likely future behavior to decide who to hold responsible, to punish, or to treat negatively.

Erosion of Trust & Big Data Backlash

The success of Big Data depends on the willingness of the public, of millions and millions of individuals individually and collectively to provide often personal data to Big Data analysts. They will do so if they see a value in letting others have that data, and if they see that others are not abusing the power that derives from having all that data. In short, Big Data success depends on user and societal trust in those that gather analyze and store data.

The moment that trust is lost, users may opt for another, less data-invasive provider. We have seen this play out in the market only recently with respect to social networking platforms. Five years ago, Facebook held a commanding lead over other platforms. Then users realized that Facebook retains all of the data, and thus creates vulnerabilities – the drunken photo from the last office party, the stupid missive angrily written then posted.

Over the last two years, alternative social networking and sharing platforms, such as *Snapchat* and *Frankly* (and many others) have cropped up and are being embraced by dozens of millions of users. *Snapchat* is said to facilitate the exchange of hundreds of millions of photos among its members every week, but these photos are ephemeral – they vanish quickly and automatically. Users have deliberately chosen *Snapchat* over Facebook because they trust *Snapchat*, but they do no longer entrust Facebook with their personal data. Frankly commits to total securing of a text while it is among sender and recipients (even *Frankly* can't read it) and completely erasing the text when done.

If Big Data users continue to gather data and extract value without keeping user trust in mind, they will destroy trust, lose customers by the millions, and end up as failures. But more is at stake: if people lose trust in sharing data with Big Data companies, the entire Big Data ecosystem may be in danger. Trust is something that can be destroyed quickly, but it takes a very long time to rebuild it if at all. Look no further that what has happened to the large US retailer, Target, following its data breach late last year. The effort to restore their customer's trust had been costly and uncertain. For a side reference, a list of the 15 worst data breaches between 2001 and 2012 has been included in Annex 1. As recent headlines are calling out to very large breaches occurred in the last six months: Target, a large U.S. retailer, compromised the personal information of somewhere between 70 and 110 million of its customers; and more recently eBay's, an online e-Commerce site, personal information for about 140 million of its customers was hacked.

Thus, it is in the self-interest of Big Data companies (and government agencies) to handle personal data with responsibility and care, and to maintain and enlarge the trust users have in their handling the data. And governments and society has an interest in ensuring that the regulatory framework is in place that helps further such trust, so that Big Data can flourish without exposing unnecessarily millions to Big Data's Dark Sides.

To this end, high-level expert groups have recently produced white papers and other documents, from the World Economic Forum and the European Union to (perhaps most recently) the White House. We have earlier mentioned the ITU's Tech Watch Report that was published late last year. While the ideas and suggestions in these efforts are varied and heterogeneous, a few trend lines emerge which we will discuss in the fourth and final section of this document.

But however these trends ultimately settle into concrete regulatory policies, they will likely require compliance to new and stricter regimes, and thus increase associated cost. While this is intended to maintain and improve user trust – essential for long-term success of Big Data – many businesses may perceive these additional costs as a negative aspect of Big Data.

False confidence in Data

Connected to, but broader than these Dark Sides that affect the acceptance of Big Data analyses in our society is another potential Dark Side that clouds the vision and understanding of those that employ Big Data. This challenge is not unique to Big Data, but Big Data is especially vulnerable to it. It is the danger that we imbue results of data analysis with more meaning that it actually has, believing that we understand more parts of reality than we actually do. It leads us to false decisions that we make with false self-confidence.

For instance, after cities introduced “predictive policing” crime decreased. Bold officials were heard suggesting that this was caused by law enforcement’s new Big Data tool. But as so often the data does not reflect causality, and thus does not prove (or even strongly suggest) that Big Data was the reason for the decline in crime. Similarly, in corporate settings marketing and advertising managers are often attributing sales successes to certain (Big Data) campaigns they ran, but without enough conclusive data to show this.

In the Big Data age we will see the world much more through a lens of data and empirics than ever before. Hopefully that improves decision-making. But it also increases the danger of falling prey to giving data more meaning than it deserves, and thus to succumb to the Siren’s Song of the confidence over data.

The Rise of the Data Barons

Finally we must also acknowledge a Dark Side of Big Data that is not directly linked to individuals and their rights and freedoms, but to data markets and the data economy. As we have mentioned combining and adding data increases its value not linearly but exponentially. This means that large data holders have a very strong incentive to grow even larger in order to extract more of the data’s intrinsic, but hidden value. Some experts fear that this may lead towards an ever-increasing concentration of data markets, stifling competition and in turn constraining innovation and destroying overall value.

For instance, Google has advanced a number of acquisitions of companies in recent years that add significantly to its ability to ingest and gather a wide variety of data. This includes its purchase of *ITA*, one of the world’s leading air travel reservation systems, or *NEST*, a company that creates devices and a platform in households to collect data about living habits – heating, cooling, even if you are at home. Similarly, Facebook has bought companies in the social networking sector to add even more data troves and users to its fold.

To an extent, this trend of concentration is countered by a lively ecosystem of Big Data startups, some of which succeed by positioning themselves well in the flow of information and compete well even against the very largest of Big Data companies. Certainly the fluidity of the Big Data ecosystem, enabled by low barriers to entry, enables these startups and act as a counterforce to market concentrations. Regulators, too, such as in the US, have attached restrictions to recent acquisitions of data companies to ensure competitive data markets.

But overall it is likely that we will have to remain vigilant against the Dark Side of market concentration and data oligopolies.

In this and the previous section we explained the upsides and the downsides of Big Data, its opportunities and its challenges. The salient question of course is whether the downsides and costs will exceed the benefits attained from the use of Big Data or vice versa. Which Big Data future is going to result? Are we going to stop Big Data, forego its many benefits in return for privacy, trust and unpredictability? Or is the pendulum swinging far into the other direction, resulting in massive Big Data uses, leaving millions of people exposed and hurting, deeply distrustful of the new technology, and creating a potent and dangerous ground for neo-luddites to fight not just Big Data but modern technology more generally? How can we negotiate a path that grants us the ability to utilize Big Data, while at the same time ensuring that Big Data is used responsibly and to the benefit not just of a handful of data holders but the wider market, and in fact society at large?

There is no simple answer to this question, but in the following, final section we aim to suggest a few policy proposals that policy makers we believe ought to consider as we approach this Big Data world, especially in the field of telecommunications.

5. The Role of (and future for) Regulatory Authorities

It is clear given the powerful qualities of Big Data and the likelihood that Big Data will shape all sectors of the economy and considering its significant dark sides, that policy makers at all levels will want to play a role on influencing Big Data's trajectory. The fundamental question however is what dimensions of Big Data policy makers should focus on in particular in their regulatory efforts. In the following we suggest four such areas of regulatory involvement:

Ensure Protection

The most obvious is of course to ensure effective protection of individuals' privacy. As we have discussed above, current mechanisms of privacy protection will become increasingly ineffective in the context of Big Data. This is not only problematic because it potentially harms affected individuals; it is also detrimental to the acceptance and utilization of Big Data, because without sufficient societal trust in Big Data users, Big Data users will not be able to operate. Thus it is not just in the interest of society, but in the very interest of all responsible users of Big Data to ensure that effective mechanisms protecting privacy are in place.

What is needed is an additional and more effective protection mechanism. Recent work undertaken by a group of privacy experts from around the world point towards a regulatory mechanism that would shift the focus of privacy protection from informed consent at the point of collecting personal data to accountable and responsible uses of personal data. The core idea is that with such a mechanism in place users of personal data would have to evaluate the privacy harms and implications of a potential new use of such data and what safeguards would need to be put in place to reduce the privacy harms before this use could commence. And while this assumes that the Big Data users would have to evaluate their intended applications, incorrect evaluations and insufficient implementation of safeguards would not only lead to civil and criminal liability. The mechanism also foresees a well-resourced privacy regulator with the expertise and power to enforce such a use based privacy protection mechanism.

The advantage of such an additional mechanism are clear: privacy protection would not rely on the mystical ability of individuals to fully comprehend the complex uses of their personal data at the moment of collection; data users could not forego the implementation of stringent privacy safeguard by pointing towards rather formal "consent of the data subjects". And enforcement would not depend on individuals suing data users (which we know from practice very, very rarely if ever happens), but rely on much more powerful privacy regulatory agencies with sufficient resources and stamina to regulate and enforce even against the most powerful data users. In return, data users would be permitted to reuse personal data for novel purposes (and thus unleash the power of Big Data) as long as a comprehensive privacy assessment had shown that it would produce minimal privacy risks.

Of course, such an additional mechanism would not solve all privacy challenges related to Big Data, but we suggest that a focus on responsible and accountable data use will go a long way in addressing some of the most troubling privacy challenges created by Big Data.

Protecting Human Volition / Taming Probabilistic Predictions

Probabilistic predictions, the operational outgrowth of Big Data analyses, can be tremendously useful. They reduce uncertainty and risk in the present and the future, and thus help Big Data users and by extension society at large to better plan and prepare for the future through better decision-making in the present. At the same token probabilistic predictions also pose unique policy challenges, especially when they are used to decide who to punish or hold responsible based only on predictions. For instance, if a government would use Big Data predictions to decide exactly which individual to put under surveillance or police heavily not because of past behavior of this individual but just because of Big Data predictions, such a policy would rightly be viewed as infringing dangerously onto human free will.

Regulatory authorities, including those intent to facilitate the use of Big Data and the growth of the data economy, are therefore well advised to put in place clear restrictions on how and for what purpose government agencies can utilize Big Data predictions. Under no circumstances can such predictions be turned into the reasons to punish people or assign individual responsibility to just forecast behavior. There must be in place a bright red line that interdicts such abuses of Big Data analysis.

Uses by government agencies as well as commercial entities of Big Data predictions of future behavior that result in negative treatment, quasi-punishment or the withholding of benefits granted to others, while not prohibited per se, must (we suggest) meet strict scrutiny. This includes providing transparency into the data analysis, as well as the guaranteed right afforded to affected individuals to disprove the prediction.

Facilitating Publicly Available Big Data Expertise

Transparency and the right to disprove predictions, as just mentioned, will only be usable for the general public, if individuals do not have to engage in confronting complex Big Data analysis themselves, but can available themselves of especially trained Big Data experts that are also obliged to help these individuals. We envision a new cadre of such experts – the "algorithmists". Specially trained, they would take vows of impartiality, confidentiality and professionalism, much like civic engineers, or doctors.

Individuals who believe they have been mistreated because of false Big Data predictions could contact algorithmists, who in turn would investigate and render a decision. They would also help individuals in disproving Big Data predictions if an individual believes such a prediction is wrong.

Algorithmists could also advise data users on how to best implement transparent, disprovable predictive decision making, and how to ensure responsibility and accountability in their Big Data predictions.

Algorithmists would have special Big Data expertise, which includes statistical and technical training, but would also be well versed in the ethical considerations at play and the legal and regulatory constraints in place.

Keeping Data Markets Fluid

So far we have focused on the role of regulatory authorities to defend and enforce the rights of the individuals in the shadow of Big Data, whether it is the right to be free from undue surveillance, unlawful use of personal data, or maltreatment based on incorrect probabilistic predictions based on Big Data analyses. But there is another, equally important dimension that is not directly related to individual rights.

As some data markets are becoming more concentrated over time, and more and more data held by fewer and fewer commercial entities, ensuring competition in the data economy becomes paramount. Otherwise Big Data may face the same fate as steel manufacturing and railways in the late nineteenth century in the US. The concentration of power of these industries in very few hands gave rise to the first effective antitrust and competition legislation in the world, and to the recognition that government plays a role in ensuring powerful, market-stifling trusts do not form, and where they have formed they are busted.

Ensuring competition in data markets can take a variety of forms. The most obvious is for data holders to be forced to let others access their data holdings under fair and reasonable terms. Such FRAND licensing (as the term of art is) has been routinely utilized in certain areas of patent protection, and shown to be effective. Moreover, the US federal government has in recent years in a number of cases already used a

FRAND[25] licensing mandate to constrain data holders power after these data holders had acquired large data sets.

The advantage of such an approach is not only that the mechanisms has already been tested and found to be effective, but that the mechanism is well known to competition authorities and thus makes it to get it employed. Moreover, such a mechanism is utilizing market competition to reduce the power of large data holders which is much preferable to more limiting restrictions or market interventions.

Some experts have gone one step further and suggested that for data markets to truly function well, one needs to put in place a legal exclusion right for data, much like we already have in place for intellectual property. Whether such a right is truly needed, and what its features and limitations would be, this paper cannot answer. It is important, however, to note these experts' opinion in this context.

6. Forums, discussions and papers

BIG – Big Data Public Private Forum

In Europe Big Data Public Private Forum (BIG) [26] is working towards the definition and implementation of a clear strategy that tackles the necessary efforts in terms of research and innovation, while also it provides a major boost for technology adoption and supporting actions for the successful implementation of the Big Data economy.

In addition each year various government and private sector entities meet to exchange their views on projects of importance in Europe. This meeting is called the European Data Forum (2014.data-forum.eu). The forum is designed to capture a larger umbrella of views by examining Open Data, Linked Data and Big Data. This year's forum included work in: open data in the transport and communications sectors in Finland; public sector information at the European Commission; the European Single Digital Market & what is required to achieve it; predicting parking supply to satisfy demand in a smart city; to name a few.

The World Economic Forum

The World Economic Forum, an international institution committed to improve the state of world through public-private cooperation, acknowledges a new approach to handle data is necessary to protect the rights and wellbeing of individuals.

One such report[27] published in 2013 carefully lays out three strong subthemes:

- From Transparency to Understanding: People need to understand how data is being collected, whether with their consent or without – through observations and tracking mechanisms given the low cost of gathering and analyzing data.
- From Passive consent to engaged Individuals: Too often the organizations collecting and using data see their role as a yes-no / on-off degree of consent. New ways are needed to allow individuals to exercise more choice and control over this data that affects their lives.
- From Black to White to Shades of Gray: the context by which data is collected and used matters significantly. How is the data used; much like money, it means little until it is used.

In order to achieve a level of trust during the flow of data at least five issues were discovered about the data: protection; accountability; empowerment; transparency and respect. There is a deep responsibility assumed for using personal data. Before the dawn of networked data, individual data was generally used

[25] A popular term for Fair, Reasonable and Non-Discriminatory terms.

[26] http://big-project.eu/

[27] WEF; Unlocking the Value of Personal Data: From Collection to Usage; http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf

once, and usually for a specific purpose. But the era of Big Data allows for analytics to reuse data to develop more value to others about that data.

In April this year, the WEF offered a report titled, *Delivering Digital Infrastructure*, Advancing the Internet Economy[28] that measurers the fast pace of technological change against the need to insure services and support infrastructure keep up. It recommends a rethink of the regulatory scope, approach and level of engagement. By scope in this age of information and speed, it recommends thinking in far broader terms – taking into account that a decision at one level impact entire economies. By approach, it touches on an oft repeated mantra; "move the ex-ante rules to ex-post, while moving the ex-post to forborne, and repeat the cycle." Finally, the report brings up the idea of level of engagement, and by that it refers to harmonization of decisions that cross national borders – specifically spectrum.

More recently the WEF released a report titled, *Risk and Responsibility in a Hyperconnected World*[29], and focuses directly on the malicious intent to disrupt or capture information (data) in both the private and public sectors.

The International Telecommunication Union

The big data approach taken by ITU so far focuses on the following areas and questions[30]. To address these increasingly important issues, reports, such as the current paper addressing the regulatory issues, are being prepared as well as workshops and dedicated sessions in ITU events.

Standardization[31]

- Which standards are required to facilitate interoperability and allow technology integration in the big data value chain?

- Which definitions, taxonomies, secure architectures and technology roadmaps need to be developed for big data analytics and technology infrastructures?

- What is the relationship between cloud computing and big data in view of security frameworks?

- Which techniques are needed for data anonymization for aggregated datasets such as mobile phone records?

- How is big data exploited in different industries; what are the specific challenges faced; and how can these challenges be addressed through international standards, e.g.,
  - **Telecommunications**: A workshop on standards for telco big data will be held on 17 June 2014 at ITU's TSAG meeting.[32]
  - **Healthcare**: Big data is a recurring theme in ITU's standardization activities on e-health.
  - **Automotive**: ITU's symposium on the *Future Networked Car*[33] highlighted the use of data analytics to making transportation safer, more efficient and more environmentally friendly.
  - **Aviation**: Following a call from Malaysia's Minister of Communications and Multimedia at WTDC'14, ITU facilitated an *Expert Dialogue on Real-time Monitoring of Flight Data, including the Black Box* on 26-27 May in Kuala Lumpur. Experts from both, the aviation and ICT sectors debated the *Need for International Standards in the Age of Cloud Computing and Big Data*,

[28] WEF; http://www3.weforum.org/docs/WEF_TC_DeliveringDigitalInfrastructure_InternetEconomy_Report_2014.pdf

[29] WEF; http://www3.weforum.org/docs/WEF_RiskResponsibility_HyperconnectedWorld_Report_2014.pdf

[30] Source: ITU Measuring the Information Society Report (2014).

[31] For further information on the work on big data carried out by the ITU Telecommunication Standardization Bureau (TSB), see http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx.

[32] http://www.itu.int/en/ITU-T/Workshops-and-Seminars/bigdata/Pages/default.aspx

[33] http://www.itu.int/en/fnc/2014/Pages/default.aspx

adopted a communiqué highlighting challenges including those specific to aviation, and proposed concrete actions for future work and standardization, in collaboration with the International Civil Aviation Authority (ICAO).[34]

Regulation[35]

- What are the key regulatory issues at stake and how can and should big data be regulated?

- How does big data impact the regulation of privacy, copyright and Intellectual property rights (IPR), transparency and digital security issues?

- What is the link between big data and open data?

- Is there a need to regulate data management and service providers?

- How can market dominance in the area of big data be prevented and the rights of the data owners protected?

ICT data collection and analysis

- How can big data complement existing ICT statistics to better monitor information society developments?

- Which type of data from ICT companies are most useful and for which purposes?

- Which new ICT indicators could be produced from big data sources?

- What are key issues that need to be addressed, and by whom, in terms of collecting and disseminating big data in telecommunications?

- What is the role of National Statistical Offices and how can big data complement official ICT data?

UNPulse

Launched by the Executive Office of the United Nations Secretary-General, to respond to the need for more timely information to track and monitor the impacts of global and local socio-economic crises, the UNPulse Initiative explores how new, digital data sources and real-time analytics technologies can help policymakers understand human well-being and emerging vulnerabilities in real-time, in order to better protect populations from shocks[36].

"The initiative was established based on a recognition that digital data offers the opportunity to gain a better understanding of changes in human well-being, and to get real-time feedback on how well policy responses are working. The overarching objective of Global Pulse is to mainstream the use of data mining and real-time data analytics into development organizations and communities of practice. To this end, Global Pulse is working to promote awareness of the opportunities Big Data presents for relief and development, forge public-private data sharing partnerships, generate high-impact analytical tools and approaches through its network of Pulse Labs, and drive broad adoption of useful innovations across the UN System"[37].

[34] http://www.itu.int/en/ITU-T/Workshops-and-Seminars/ccsg/expdial/Pages/default.aspx

[35] A background document on big data that was prepared for the GSR 2014 is available at http://www.itu.int/en/ITU-D/Conferences/GSR/Pages/gsr2014/default.aspx.

[36] Adapted from UNPulse (About) http://www.unglobalpulse.org/about-new

[37] Id.

US: White House, Big Data Initiative

Two major initiative[38] to fund research on six “Big Data” initiatives were announced by the President two year ago to:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

This year, the Executive Office released two reports: The first report[39] Big Data: *Seizing Opportunities, Preserving Values*, is a comprehensive treatment of the subject. The report urged focus that will move the privacy discussion forward by preserving Privacy Values(both in the United States and through interoperable global privacy frameworks); educating Robustly and Responsibly: Recognizing schools as an important sphere for using big data to enhance learning opportunities, while protecting personal data usage and building digital literacy and skills; Big Data and Discrimination: Preventing new modes of discrimination that some uses of big data may enable; law Enforcement and Security: ensuring big data’s responsible use in law enforcement, public safety, and national security; and harnessing data as a public resource, using it to improve the delivery of public services, and investing in research and technology that will further power the big data revolution.

The second report[40], *Big Data and Privacy: A Technology Perspective* examined the nature and evolution of technology and its capabilities and the challenges surrounding protecting individual privacy. What is useful to know from this work is that it concludes with the notion that technology alone cannot protect privacy; policy needs to play a strong role and needs to reflect what is technologically feasible.

This report’s policy position centered on five recommendations: to Focus on the use of Big Data, and less on collection; to avoid embedding technological solutions into policy, to focus on research and deployment of technologies that help to protect privacy, to encourage education and career professions and for the US to take the lead both internationally and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today.

It is worth noting that this report became subject to quick debate from privacy advocates indicating that it relies on policy aimed at the use of data, and not as much on its collection. By extension the criticism thought there would be discrimination against the poor, elderly and minorities, and even children – those not in a position to protect themselves – by placing the burden of protection on the individual.

7. The Wrap

The world of Big Data is in its infancy, taking its first steps in what will be a long journey. It will be guided to an extent by decisions made by the Regulatory Authorities in regions and jurisdictions throughout the world. The options are varied, complex, and risky.

We must not lose sight of its great potential, benefitting the individual, organizations and society as a whole. We explained in this paper the underpinnings of why we are at the crossroads of Big Data; what

[38] http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

[39] Executive Office of the President; May 1, 2014; http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[40] Executive Office of the President; May 1, 2014; http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

factors have put us in a position to see this opportunity. From those points we see that there is no evidence to believe the trends will reverse anytime soon. We shared only a small handful of benefits already realized by the promising use of Big Data.

We also know that there are a growing number of concerns to protect not only the interests of the individual, but the ability to innovate. This will be the balance struck by those in a position guide or control these opportunities. The public will continually have to believe that there is greater benefit than cost to them to avoid backlash or loss of trust. Such a loss would promise a reversal of the gains seen. What should the Authorities do? We have condensed our policy focus to four points:

- Ensure protection not only for society itself, but for those users of Big Data. Recognize the shift to protect at the point of use and away from the point of collection.
- Protect human free will. Predictive approach to determining societies decisions must be carefully managed to avoid hold those accountable only based on prediction.
- Grow the skills pool of people capable to manage this properly. Talent will be required and needed to insure we understand what we are doing but to give those who believe they are aggrieved by the consequences of Big Data, and are in a position for proper redress.
- Keep data market fluid through a number of oversight tools. Data markets must be kept fluid and robust and proper frameworks are needed to insure that small groups or individuals become the earlier monopoly Trusts that controlled the health and degree of innovation in a segment of the economy.

With the right level of attention today, our children and their children will find a world that has benefited from the creativity and imagination that Big Data offers.

Annex 1 – 15 Worst Data Breaches (2000-2012)[41]

1. Heartland Payment Systems

Date: March 2008

Impact: 134 million credit cards exposed through SQL injection to install spyware on Heartland's data systems. The vulnerability to SQL injection was well understood and security analysts had warned retailers about it for several years. Yet, the continuing vulnerability of many Web-f acing applications made SQL injection the most common form of attack against Web sites at the time.

2. TJX Companies Inc.

Date: December 2006

Impact: 94 million credit cards exposed.

There are conflicting accounts about how this happened. One supposes that a group of hackers took advantage of a weak data encryption system and stole credit card data during a wireless transfer between two Marshall's stores in Miami, Fla. The other has them breaking into the TJX network through in-store kiosks that allowed people to apply f or jobs electronically.

Date: March 2011

Impact: Exposed names and e-mails of millions of customers stored in more than 108 retail stores plus several huge financial firms like CitiGroup Inc. and the non-profit educational organization, College Board. The source of the breach is still undetermined, but tech experts say it could lead to numerous phishing scams and countless identity theft claims. There are different views on how damaging the Epsilon breach was. Since Epsilon has a client list of more than 2,200 global brands and handles more than 40 billion e-mails annually, it could be, "the biggest, if not the most expensive, security breach of all-time."

4. RSA Security

Date: March 2011

Impact: Possibly 40 million employee records stolen.

[41] Taylor Armerding, CSO Online; 15Feb12, 15 Worst data breaches of the 21st Century

The impact of the cyber-attack that stole information on the company's SecurID authentication tokens is still being debated. The company said two separate hacker groups worked in collaboration with a foreign government to launch a series of spear phishing attacks against RSA employees, posing as people the employees trusted, to penetrate the company's network. Among the lessons are that even good security companies like RSA are not immune to being hacked. Finally, "human beings are, indeed, the weakest link in the chain".

5. Stuxnet

Date: Sometime in 2010, but origins date to 2007

Impact: Meant to attack Iran's nuclear power program, but will also serve as a template f or real-world intrusion and service disruption of power grids, water supplies or public transportation systems. The immediate effects of Stuxnet were minimal -- at least in this country -- but it ranks it among the top large-scale breaches because, "it was the first that bridged the virtual and real worlds. When a piece of code can have a tangible effect on a nation, city or person, then we've truly arrived in a strange, new world," he says.

6. Department of Veterans Affairs

Date: May 2006

Impact: An unencrypted national database with names, Social Security numbers, dates of births, and some disability ratings for 26.5 million veterans, active-duty military personnel and spouses was stolen. The breach pointed once again to the human element being the weakest link in the security chain. The database was on a laptop and external hard drive that were both stolen in a burglary from a VA analyst's home. The analyst reported the May 3, 2006 theft to the police immediately, but senior officials at the Veterans Affairs were not told of it until May 16. The VA estimated it would cost $100 million to $500 million to prevent and cover possible losses from the theft.

7. Sony's PlayStation Network

Date: April 20, 2011

Impact: 77 million PlayStation Network accounts hacked; Sony is said to have lost millions while the site was down f or a month. This is viewed as the worst gaming community data breach of all-time. Of more than 77 million accounts affected, 12 million had unencrypted credit card numbers. According to Sony it still has not found the source of the hack. They gained access to full names, passwords, e-mails, home addresses, purchase history, credit card numbers, and PSN/Qriocity logins and passwords.

8. ESTsoft

Date: July-August 2011

Impact: The personal information of 35 million South Koreans was exposed after hackers breached the security of a popular software provider. It is called South Korea's biggest theft of information in history, affecting a majority of the population. South Korean news outlets reported that attackers with Chinese IP addresses uploaded malware to a server used to update ESTsoft's ALZip compression application. Attackers were able to steal the names, user IDs, hashed passwords, birthdates, genders, telephone numbers, and street and email addresses contained in a database connected to the same network.

9. Gawker Media

Date: December 2010

Impact: Compromised e-mail addresses and passwords of about 1.3 million commenters on popular blogs like Lifehacker, Gizmodo, and Jezebel, plus the theft of the source code for Gawker's custom-built content management system. Online forums and blogs are among the most popular targets of hackers.

10. Google/other Silicon Valley companies

Date: Mid-2009

Impact: Stolen intellectual property

In an act of industrial espionage, the Chinese government launched a massive and unprecedented attack on Google, Yahoo, and dozens of other Silicon Valley companies. The Chinese hackers exploited a weakness in an old version of Internet Explorer to gain access to Google's internal network. It was first announced that China was trying to gather information on Chinese human rights activists.

11. VeriSign

Date: Throughout 2010

Impact: Undisclosed information stolen

Security experts are unanimous in saying that the most troubling thing about the VeriSign breach, or breaches, in which hackers gained access to privileged systems and information, is the way the company handled it -- poorly. VeriSign never announced the attacks. The incidents did not become public until 2011, through a new SEC-mandated filing. VeriSign said no critical systems such as the DNS servers or the certificate servers were compromised, but did say that, "access was gained to information on a small portion of our computers and servers."

12. CardSystems Solutions

Date: June 2005

Impact: 40 million credit card accounts exposed.

CSS, one of the top payment processors for Visa, MasterCard, and American Express is ultimately forced into acquisition. Hackers broke into CardSystems' database using an SQL Trojan attack, which inserted code into the database via the browser page every four days, placing data into a zip file and sending it back through an FTP. Since the company never encrypted users' personal information, hackers gained access to names, accounts numbers, and verification codes to more than 40 million card holders.

13. AOL

Date: August 6, 2006

Impact: Data on more than 20 million web inquiries, from more than 650,000 users, including shopping and banking data were posted publicly on a web site. In January 2007, Business 2.0 Magazine ranked the release of the search data in among the "101 Dumbest Moments in Business." AOL Research released a compressed text file on one of its websites containing 20 million search keywords for more than 650,000 users over a three-month period. While it was intended for research purposes, it was mistakenly posted publicly. AOL pulled the file from public access by the next day, but not before it had been mirrored and distributed on the Internet.

14. Monster.com

Date: August 2007

Impact: Confidential information of 1.3 million job seekers stolen and used in a phishing scam. Hackers broke into the U.S. online recruitment site's password-protected resume library using credentials that Monster Worldwide Inc. said were stolen from its clients.

15. Fidelity National Information Services

Date: July 2007

Impact: An employee of FIS subsidiary Certegy Check Services stole 3.2 million customer records including credit card, banking and personal information. Network World reported that the theft was discovered in May 2007. But the theft was not disclosed until July. An employee allegedly sold the data for an undisclosed amount to a data broker.