Chapter 9
Data-Intensive Computing and the Future of Research

Åke Edlund

Abstract Massive data sets are being produced in industry and academia of today. Scientists are probing extreme phenomena in scientific fields with mature theories like astrophysics and particle physics. At the same time we see increasingly exploratory research areas evolve, mining large data sets to find new phenomena and patterns. In industry, but also very much in academia, there are huge efforts in making meaning of human activity on the Internet, and as if these data sets were not enough, sensor networks 'sensing everything everywhere' is evolving. Information advantage, be it in business or academia, is crucial in today's global competition, and that is why there is so much interest in data and the technologies handling the data. What is new in the discussions about data and its underlying value is the increasing rate in the production of information, and how companies and academia are cross-fertilizing the information flows to produce even more information. Internet, Cloud Computing, 'Big Data', Internet of Things – it is easy to get lost in the technical discussions forgetting what it is all about: information, how to gather it, how to manage it, and how to make timely and informed decisions based on what we find. During the last decade much of the discussions have been centered on the effects of the cloud computing paradigm shift, but that is only the latest technological achievements in the overall effort of producing and analyzing information. In this chapter we look into the characteristics and evolution of information technology, discussing in more detail the latest paradigm shifts, and the new challenges and opportunities facing the companies and scientists. In the end of the chapter we include a list of suggested research topics in this area.

A. Edlund (✉)
KTH Royal Institute of Technology, Stockholm SE-100 44, Sweden
e-mail: edlund@pdc.kth.se

E. Ullberg (ed.), *New Perspectives on Internationalization and Competitiveness*, DOI 10.1007/978-3-319-11979-3_9

Software Industry Characteristics and Its Paradigm Shifts

Since the move into computing, the evolution of Science is closely intertwined with software industry, adopting to its changes but also directly affecting the software industry itself. Below we briefly look at the characteristics of software industry and how, and why, it is evolving at an accelerated pace.

The Power of Exponential Changes

Exponential improvements in computer hardware over the last decades have propelled the software industry with wide implications in all information centric areas. Computers are able to perform ever increasingly number of operations per second, doubling in every 18 months for the same cost unit (Moore's law), and in parallel storage capabilities are improving, even if not at the same speed. To further illustrate the exponential changes, we have added the picture below showing the, even higher, increase of capabilities from the area of genome sequencing. The left picture gives the reader a hint of the upcoming data flood from the Life Sciences area, both from industry as from academic research. The right picture shows the data explosion following the change. Just to clarify: sequencing a genome cost years of work and hundreds of MUSD in the beginning of the century. Now the cost is down to below 1,000 USD (www.utsandiego.com/news/2014/Jan/14/illumina-thousand-dollar-genome; Haussler; Haussler et al. 2012), and sequencing time to minutes (Figs. 9.1 and 9.2).

In addition to this technological evolution, we have seen the rollout of high-speed network connections and an enormous increase of Internet users. We have also seen the change in how users connect to the Internet: The selected means of accessing the services from Internet has moved the market from PCs to handheld

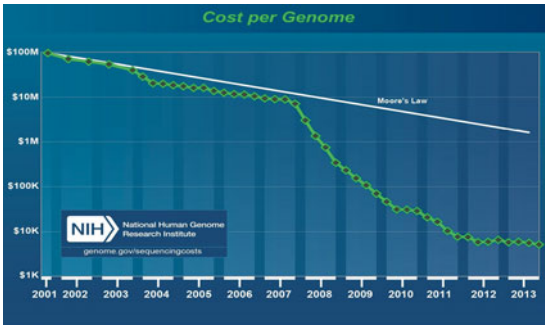

Fig. 9.1 (www.genome.gov/sequencingcosts)

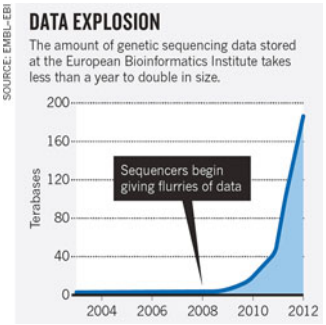

Fig. 9.2 (EMBL-EBI)

devices, with implications on networks, businesses in general and in challenging the earlier Windows-Intel dominance: the users of handheld clients need different solutions both in terms of software as in technological architecture. We also see how developing countries 'leap frog' over the old wired investments and move straight to the mobile Internet business models.

In parallel with the technological changes, its users are adopting to the new services. This change in user's behavior, e.g. in accepting payments over the Internet, enables new business models to emerge. Cloud computing, discussed more in depth later in this chapter, relies heavily on these changes in user behavior. Trust is a key component in this user behavior. It took years for the industry to earn its customers' trust. But trust is lost faster than it is earned.

Adopting to this change in user behavior and technology a new set of dominant companies that didn't even exist two decades ago have evolved, challenging the older giants. So, what is the main difference between the old way (before Internet) of doing business to todays? The short answer is that the new dominant companies are fully information focused, they deliver services over the Internet, and they have been collecting information from start, both to be able to improve their own services as well as creating new services based on the enormous amount of data they have about the Internet. In short, one of Google's main ideas is to "download the Internet and rank it". This is the way they order their data for us to be able to use e.g. Google search. But, they also store what we search for, what we click on, and from where we did this. Based on this information Google analyze their data, combines the data sets to create new, improved, services to its user. Other examples are the media streaming companies, who are very active in building recommendation systems on music and films to help improving their user's experience. At the same time they gather geographical and behavior information to adjust their underlying delivery technology – again to improve their user's experience, but also to lower their own cost.

The more advanced data analysis the companies can manage, the better they will cope with competition. This is 'Big Data', and, yes, the data the larger companies are analyzing is big in volume, as in velocity (how much it increases over time) and variety (very different sources with high variation in quality). But is this only for larger companies? No. Any company who deliver services over the Internet have large amount of user and usage data, data that hides pattern and insights on how to improve – how to compete.

Software Industry Characteristics and Its Paradigm Shifts

> Unlike hardware, software is expected to grow and evolve over time. Whereas hardware designs must be declared finished before they can be manufactured and shipped, initial software designs can easily be shipped and later upgraded over time. Basically, the cost of upgrade in the field is astronomical for hardware and affordable for software. (Patterson and Fox 2012)

The software industry is one of the most rapidly changing areas in the economy, and the software industry today is affecting most areas using information in any format. Cloud computing is the latest big change, affecting the way we produce and consume software products and services. This change is most likely greater than the introduction of Internet. The cloud market is global and it is all about services consumed over the Internet directly by customers.

Before going any further in the discussion, let's look at some numbers on why software is a very important part of industry, taking Sweden (Report from Swedish software organization Swedsoft) as an example: At Ericsson 80 % of their investments in R&D are software related – a total of 3 billion USD every year. Maybe more surprising, are the numbers from the car industry indicating that 25–35 % of the value of a car is in its software. Thirty years ago this number was 1 %. Seventy percent of the innovation built into Swedish trucks today comes from software developed in-house. Even industries closer to hardware rely heavily on software in maintaining a high productivity and competitiveness on the global market.

Software Industry Is in Constant Change

While Internet created a new way of communicating data between users and companies, Cloud Computing paves the way for a service based economy – where customers consume services – not just data – online. Instead of buying, installing and managing programs on your computer to handle your business, you go online to manage and use all your services. There is no need to handle versions of software, security patches and hardware. All you need is an Internet connection and a device

to access your services. As a reflection of this change, and as mentioned earlier, we are now moving into the 'post PC era', where smart clients (mobiles, tablets) are good enough to solve many of our daily needs.

Factors explaining this rapid development in the software industry can be found in the fundamentals of software itself: new software development is based on old – successful – software development. That is: the longer this field evolves, the more it is building tools to create new software, solving more complex problems in a shorter time. This is true for all fields, but in software the change is very rapid, as is the uptake and the inheritance (and copying) of previous results. Another fundamental characteristic of software is how easily the resulting product – the software – is duplicated and distributed. Compared to classic industry products, for example cars, software evolves and spreads considerably faster. Moreover, as was mentioned in the introductory quote "the cost of upgrade in the field is astronomical for hardware and affordable for software", further emphasizes the differences between hardware and software. Due to this feature, i.e. that software is undergoing constant change and continuous updates; software products can have very long life times (Patterson and Fox 2012).

With this in mind it comes as no surprise that we, again, face a large transition in software industry, an industry where paradigm shifts seems to appear with a regularity of once every decade.

The exponential growth in the underlying capabilities of the hardware delivering the software based applications have taken us from local computers, available only for a few national institutions, to personal mobile handheld clients accessing services where ever we go: Internet connects us, the Cloud deliver the services, and now we increase our gathering and analysis of the data surrounding us. This, latest, step is named 'Big Data' and is as disparate as 'Cloud computing' was still in 2007, and, as in the advent of Cloud computing, many consider it as no change but just something we have been doing all the time. And, yes, analyzing large data sets to gain competitive advantage is not new to larger corporations. What is new is the amount of available data and the increased capabilities to analyze the data. As described above, software industry builds on the shoulders of earlier achievements. A comparison to illustrate this: Old software licensing model, customer buys software to be used locally. The selling company receives information during the purchase and when the customer downloads updates of the software. Companies who are information centric and deliver their services over the Internet receive a flow of user information for the full duration of the usage of the service. For example, the software-as-a-service company immediately sees when the usage of the service drops – a signal that the customer might be unsatisfied with the service. This applies also to smaller companies, e.g. game developing companies who analyze in detail the usage of their games, looking for improvements to e.g. avoid making too hard (or too easy) steps in the games. This latter illustration is a good example of analyzing the company's internal data as well as public Internet data, where they look for increase in usage of 'cracking' solutions – where the game players got stuck and look for tricks to get passed the game steps.

Making Meaning Out of Massive Data: Inference Challenges

In earlier sections we discussed the characteristics of software industry. This was needed to understand why we see yet a new change in this area, just as we start to adapt to the latest paradigm shift. We write 'Big Data' of two reasons. Firstly, 'Big Data' is a very vague description of a huge area, an area that doesn't easily describe itself in just two words. Secondly, what is 'Big' in 'Data'? Learning from recent history: in the beginning of Cloud computing there were numerous definitions of cloud computing. After the overview report (Armbrust et al. 2009) the area cleared up, followed by the definitions NIST (US National Institute of Standards and Technology) and EU (2012) realizing that the meaning of the words 'Cloud computing' should be divided in some key concepts together with user dependent views. In [Michael Jordan] efforts are put into identifying the challenges in a scalable way, defining the fundamental questions regardless of what we consider 'big' today. Remember when 1 GB was huge? Not anymore.

The Long Tail of Science
Collectively "long tail" science is generating a lot of data, estimated at over 1PB per year and it is growing fast. 80-20 rule: 20 % users generate 80 % data but not necessarily 80 % knowledge. (Gannon).

Inference of Massive Data

Extracting inference out of massive data sets is challenging, creating a demand of very special combination of knowledge: understanding of the underlying data, combined with computer skills and rigorous mathematical and statistical background (Fig. 9.3).

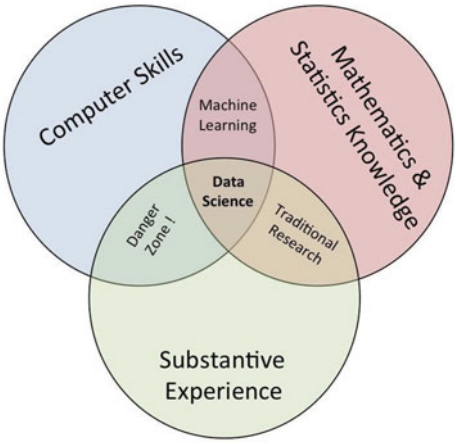

Fig. 9.3 (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram)

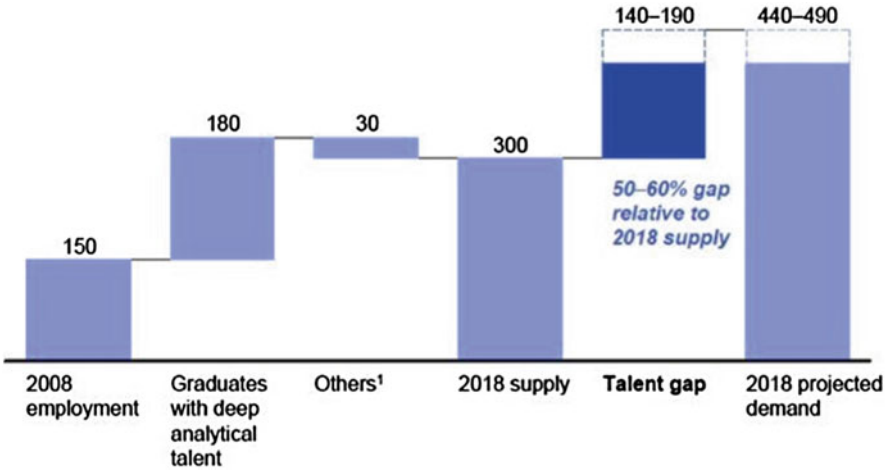

Fig. 9.4 (www.mckinsey.com/mgi/publications/big_data/index.asp)

Failing on one of these capabilities result in increased risk of mistakes, in risk of misleading conclusions. "There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions. Informatics aimed at 1.5 million jobs. Computer Science covers the 140,000 to 190,000" (www.mckinsey.com/mgi/publications/big_data/index.asp) (Fig. 9.4).

Challenges: Separating Knowledge and Misleading Information

A major challenge in analyzing massive data forces us to add even more workload: adding analyses of the error bars, the confidence interval, to the overall analysis. The challenge lies in parallelizing the error estimation, an area still in its infancy much as a result of its inherent complexity combined with the massive size of the underlying data. Or as Michael I Jordan (Frontiers in Massive Data Analysis report) describes it: "There is a need of statistical principles (that scale) to justify the inferential leap from data to knowledge. It is always possible to turn data into something resembling knowledge but which actually is not. And it can be quite difficult to know that this have happened".

Data Exploration: Not So Easy

Exploring these massive data sets to infer knowledge is a demanding task also with today's data-intensive technologies. In parallel with the increased performance of base technologies, mathematical algorithms are evolving with a multitude of challenges in developing traditional statistical methods. Machine learning, a sub discipline of statistics, gives the analysts novel approaches to classify and identify patters in the underlying, massive, data sets. Still, with all these improvements in methods and infrastructure, the massive data sets creates challenges to the user, also in the choices on how much information she should ask for. It is easy to become greedy with such wealth of information.

Not knowing the underlying laws of large datasets puts the user at risk of misleading conclusions. The problem arises when the user adds more features to be studied, increasing the possible correlations exponentially. Adding more features to the models increase the risk of 'perfect matching' of features that only share probability distributions – not having anything else in common. We find inference, that doesn't have bearing in reality, and the bad news is: we won't even understand, neither notice, the mistake.

Time Aspect of Inference

As mentioned above there is a risk of becoming greedy when dealing with large datasets. Adding too many features into the equation is not the only sign of greed here: what if I can get the conclusions, the recommendations, from the data sets faster? Faster than anyone else? We see an increasing interest in data analytics in near real-time, going from reporting to operations. Moving the data analysis from batch processing, longer analyzing time spans, towards analyzing near-time to real-time processing of information increase the challenges we already discussed. If we try to analyze streaming data we have a shorter time window to do the actual analysis, limiting us further. If you have the required knowledge in statistics you will adapt the questions you pose to such data.

Data Discovery and the Internationalization of Science

Science does not evolve in paradigm shifts as frequently as software industry; still we see three clear changes in Science in the history (see The fourth paradigm: Data-intensive scientific discovery). In the very beginning science was an empirical discipline, describing natural phenomena. Over time, based on the patterns we identified, we started to build theories describing many of the phenomena, using models and generalizations. This was the first paradigm shift, expanding from empirical to theoretical science. In the last few decades we have increasingly used

computers to simulate complex phenomena where analytical solutions of the theoretical models have been too, often impossible, to handle. This was the second paradigm shift in Science, adding computers to the chain of scientific work, from empirical studies to theoretical models to be translated to computers for massive simulations to further extend the reach of Science. As a result of this 'coupling' with computers and Science, these to areas are now evolving in symbiosis.

Massive data is not only generated from social interactions on the Internet and other web based information, but as much about data generated and analyzed in natural sciences, economy to humanities. Here the scientists are seeing a 4th paradigm in Science itself – the data exploration era.

Now we see a fourth paradigm shift in Science (The fourth paradigm: Data-intensive scientific discovery), data exploration, where the scientists analyze massive data sets from simulations and experiments to infer new knowledge or verify theories. The world of science has changed, where the new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, *fourth paradigm* for scientific exploration (The fourth paradigm: Data-intensive scientific discovery), often named as eScience.

Looking at the various scientific areas we see the evolution of two branches of every discipline. For example (The fourth paradigm: Data-intensive scientific discovery), if you look at ecology, there is now both computational ecology, which is to do with simulating ecologies, and eco-informatics, which is to do with collecting and analyzing ecological information. Similarly, there is bioinformatics, which collects and analyzes information from many different experiments, and there is computational biology, which simulates how biological systems work and the metabolic pathways or the behavior of a cell or the way a protein is built.

Geoffrey Fox (Hey et al. 2012) has described this change in Science in the following "Big Data Ecosystem in One Sentence"

Use Clouds running Data Analytics processing Big Data to solve problems in X-Informatics (or e-X)

X = Astronomy, Biology, Biomedicine, Business, Chemistry, Crisis, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly Spans Industry and Science (research)

With respect to internationalization of Science, eScience is taking one step further from earlier, highly international, environments. Data from large experimental devices, e.g. telescopes and particle accelerators like the CERN Large Hadron Collider (LHC), is distributed to scientists world-wide to study. Sometimes this comes as a necessity, due to the data sizes (like in the LHC case), where the data

need to be analyzed over a large number of collaborating scientists. CERN LHC is in itself a good example of international collaboration, where countries goes together to build a experiment too expensive for any single country, then sharing the data for remote analysis. It is also a good example of massive data exploration. Finding the 'needle in the haystack', the Higgs Boson, was just the beginning.

Since its move into computing many science areas are taking advantage of the paradigm shifts in software industry and is often one of the contributors, e.g. in the designing of the world-wide-web. Cloud computing, briefly mentioned above, is embraced by the science community creating new scientific services simplifying the daily work of the researchers. As seen from the Geoffrey Fox quote above, the examples are many where researchers reach out to cloud resources for increasingly larger computational challenges. Platforms for sharing and further developing workflows and data are today common practice in many disciplines, e.g. within the Galaxy community (galaxyproject.org). In the same way as enterprises, especially early stage companies, adapt to web based collaboration and communication, researchers are getting increasingly used to the same tools, e.g. using various web based project collaboration tools and code sharing like github. Today's researcher gets far without having to buy and manage hardware. For data scientists, the following section is possibly the largest change so far.

Data Analysis: As a Service

One of today's barriers to a successful data analysis environment, where the scientist can focus on the analysis on his data, is the management of the underlying infrastructure and workflows. Even if the emerging cloud technologies are simplifying the management of the infrastructure and development of the services needed, it is still a complex and demanding task. Infrastructure-as-a-Service (IaaS) gives the user elastic and cost efficient usage of the infrastructure (compute, network and data storage), while Platform-as-a-Service (PaaS) gives the user tools to develop the final services to be used. So, even if IaaS and PaaS are simplifying the basic infrastructure and development, the data analytics stacks needed (consisting of many layers and complex workflows) creates an overall complex environment for the data analyst to handle.

A number of companies are providing services to simplify the deployment and handling of these data analysis environment, from software distribution companies to larger cloud providers. Still much of the work remains for the user, and in addition the data analytics workflows themselves are often combinations of many different services (e.g. streaming, batch, graph data analysis, machine learning algorithms) with need of reloading of data. There is a need of a unifying data analytics stack and one of the most promising is the Berkeley Data Analytics Stack (BDAS) based on the Apache Spark (Zaharia 2014). In BDAS the services all use the same underlying data abstraction enabling the user to write complex analysis within one unified workflow.

The final step, in the simplification of the data analyst's work, was recently taken by one of the founding companies behind the BDAS, in *delivering the whole data analytics stack as a service, including the underlying infrastructure needed* (http://databricks.com/cloud/). By this the user now can manage and analyze his data from a browser, with a minimum need of management of the underlying infrastructure.

The implications of this new move, to a Data-Analysis-as-a-Service, are many. The amount of researcher that will now be able to do more analysis will increase dramatically. The analysts will be able to use more advanced workflows, and handle larger amount of data.

The challenge lies, as before with IaaS, the concerns with respect to data privacy: the above-mentioned Data-Analysis-as-a-Service relies on cloud providers, and at this stage only US-based, starting with Amazon – later Google and Microsoft.

Research Topic Data-Analysis-as-a-Service, as provided in (http://databricks.com/cloud/)*, gives researchers a considerably improved environment for their work. It also enables a larger set of researchers to do more science than before. Will this be a competitive advantage for the US-based researchers? Will non-US researchers be limited by their government in how much they may use the US-based infrastructures behind the Data-Analysis-as-a-Service? What will the implications be, and will this lead to a push on non-US-based infrastructure providers to evolve?*

Concluding Summary: Data Discovery and the Internationalization of Science

The academic community is increasingly making use of the same software technologies as industry. In many areas researchers in academia are early adopter of the new technologies and often part of its development. We have seen this during the introduction of Internet, where e.g. the development and specifications by Tim Berners-Lee of HTML were made due to needs of researchers at CERN. The adoption of cloud technologies by academia is well described in the XSEDE report (2013) presenting data (2012–2013) from 80 cloud users (world-wide) and their experiences. In the era of massive data analysis ('Big Data') we see an increasing contribution to the open source with novel data analytics stacks (e.g. The Berkeley Data Analytics Stack, partially hosted under Apache), and with new services emerging (as described in the section "Data Analysis: As a Service" section above).

The dependencies on the novel technologies are much the same in academic research as in industry. For example academic research is sensitive to security issues and levels of trust much in the same level as companies, especially in the areas where sensitive personal data might be affected. In Life Sciences there are limitations on how much researchers are allowed to use public clouds.

One research topic in this area is to study *how fragile current business models are, including academic users? What are the effects of losing trust in e.g. US based companies due to the news regarding NSA and personal information? What is the*

geographical and political distribution of these effects? How does it affect current public vs private cloud computing services? See e.g. "How the NSA Almost Killed the Internet" (Wired, Jan. 2014)

As was described in this chapter, handing massive data is challenging. To be able to extract knowledge out of the data, maintaining a rigorous measure of the error rate of the hypothesis made out of the data, calls for expertise in multiple disciplines. One possible research topic could be to further study the following question:

Are we getting more informed or are we just increasing the amount of misleading 'statistical' advice? How well are the analysis performed, are we sacrificing error estimates calculations for more complex analysis – or put differently, how do we strike the balance of how much we try to analyze with solid statistical handling of the information? It is hard to handle massive data, even harder if we want to have estimates of the error in our results.

Worth noting in this are is that we didn't even mention the other challenges in managing massive data sets: how to handle the increasing inflow of data, how to clean the data, how to store (and decide what to not store). Another main topic to address is the value of the analysis put in perspective of the management cost of the data handling. This is related to the question of competitive advantage in analyzing existing data and the risk in not analyzing.

Relevant to all above is the question of available persons with the right knowledge, a research topic in its own right:

If the demand is higher than the available resources, how is the market evolving for data analysts? How are the companies and nations competing for the persons with data analysis profiles – and how are data analysis companies evolving to create businesses in this gap?

One more area to study arises when considering the following studies from an academic viewpoint: "Computing is being transformed, new companies are emerging. Many organizations that have Big Data don't have the ability to process Big Data." from the Best Practices in Big Data Storage report, Tabor Communications Custom Publishing Group.

From the report "From Value to Vision: Reimagining the Possible with Data Analytics, What makes companies that are great at analytics different from everyone else" by MIT Sloan Management Review and SAS Institute (www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/reimagining-possible-data-analytics-106272.pdf) we see an increase in number of companies that see business analytics as a competitive advantage, rising from 37 % in USA in 2010 to 67 % in 2013. In the report they also estimate the number of innovative companies (in making use of business analytics) to 11 % while 29 % still remain 'analytically challenged, where the available data is more of a burden than an asset (Figs. 9.5 and 9.6).

Research Topic A similar study of the academic researchers would be very interesting, knowing that not all are early adopter of new technology.

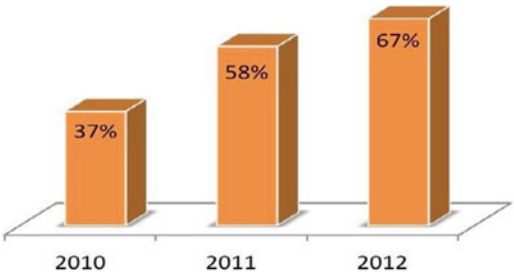

Fig. 9.5 (www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/reimagining-possible-data-analytics-106272.pdf)

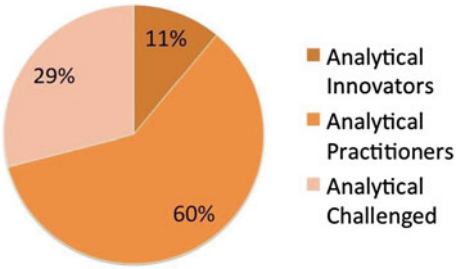

Fig. 9.6 (www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/reimagining-possible-data-analytics-106272.pdf)

We end this discussion chapter with a positive note from the central report *Frontiers in Massive Data Analysis* (Frontiers in Massive Data Analysis report) "The hope is that if massive data could be exploited effectively, science would extend its reach, and technology would become more adaptive, personalized and robust". Challenges aside, there is considerable value to find in the data deluge we are now experiencing – in all information centric research areas, i.e. basically all science.

References

Armbrust, M. et al. (2009). Technical Report No. UCB/EECS-2009-28 http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

EMBL-EBI. www.genome.gov/sequencingcosts

From advances in clouds, EU Cloud Expert Group, May 2012.

Frontiers in Massive Data Analysis report. www.nap.edu/catalog.php?record_id=18374

Gannon, D. (Microsoft Research). www.microsoft.com/eu/transforming-business/article/cloud-policy-big-data-and-the-long-tail-of-science.aspx

Haussler, D. *Today is the dawn of personal genomics*. USCS, http://www.genome.gov/sequencingcosts/

Haussler, D. et al. (2012). *A million cancer genome warehouse*. http://rr.soe.ucsc.edu/sites/default/files/rrd-2012-haussler.pdf

Hey, T., Gannon, D., & Pinkelman, J. (2012). The future of data-intensive science. *IEEE Computer, 45*(5), 81–82.

Patterson, D. A., & Fox, A. (2012). *Engineering long-lasting software: An agile approach using SaaS and cloud computing (Alpha edition)*, 2012. Available in electronic format only (e.g. Kindle from Amazon, or iBook from Apple).

Report from Swedish software organization Swedsoft. www.swedsoft.se

The fourth paradigm: Data-intensive scientific discovery. research.microsoft.com/en-us/collaboration/fourthparadigm/

US National Institute of Standards and Technology. http://www.nist.gov/, http://www.nist.gov/itl/cloud

XSEDE Cloud Survey Report, Sept 2013. www.xsede.org/xsede-nsf-release-cloud-survey-report

Zaharia, M. (2014). *An architecture for fast and general data processing on large clusters*. Technical Report No. UCB/EECS-2014-12, 3 Feb 2014. http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-12.html