Data Scientist: The Engineer of the Future

Wil M.P. van der Aalst

Scientific Director of the Data Science Center Eindhoven (DSC/e),
Eindhoven University of Technology, Eindhoven, The Netherlands.
w.m.p.v.d.aalst@tue.nl

Abstract. Although our capabilities to store and process data have been increasing exponentially since the 1960-ties, suddenly many organizations realize that survival is not possible without exploiting available data intelligently. Out of the blue, "Big Data" has become a topic in board-level discussions. The abundance of data will change many jobs across all industries. Moreover, also scientific research is becoming more data-driven. Therefore, we reflect on the emerging *data science* discipline. Just like computer science emerged as a new discipline from mathematics when computers became abundantly available, we now see the birth of data science as a new discipline driven by the torrents of data available today. We believe that the *data scientist* will be the engineer of the future. Therefore, Eindhoven University of Technology (TU/e) established the Data Science Center Eindhoven (DSC/e). This article discusses the data science discipline and motivates its importance.

Key words: Data Science, Big Data, Process Mining, Data Mining, Visual Analytics, Internet of Things

1 Always On: Anything, Anytime, Anywhere

As described in [9], society shifted from being predominantly "analog" to "digital" in just a few years. This has had an incredible impact on the way we do business and communicate [12]. Society, organizations, and people are "Always On". Data is collected *about anything*, *at any time*, and *at any place*. Gartner uses the phrase "The Nexus of Forces" to refer to the convergence and mutual reinforcement of four interdependent trends: social, mobile, cloud, and information [10]. The term "Big Data" is often used to refer to the incredible growth of data in recent years. However, the ultimate goal is not to collect more data, but to turn data into real value. This means that data should be used to improve existing products, processes and services, or enable new ones. *Event data* are the most important source of information. Events may take place inside a machine (e.g., an X-ray machine or baggage handling system), inside an enterprise information system (e.g., a order placed by a customer), inside a hospital (e.g., the analysis of a blood sample), inside a social network (e.g., exchanging e-mails or twitter messages), inside a transportation system (e.g., checking in, buying a ticket, or passing through a toll booth), etc. Events may be "life events", "ma-

chine events", or both. We use the term the *Internet of Events* (IoE) to refer to all event data available. The IoE is composed of:

- The *Internet of Content* (IoC): all information created by humans to increase knowledge on particular subjects. The IoC includes traditional web pages, articles, encyclopedia like Wikipedia, YouTube, e-books, newsfeeds, etc.
- The *Internet of People* (IoP): all data related to social interaction. The IoP includes e-mail, facebook, twitter, forums, LinkedIn, etc.
- The *Internet of Things* (IoT): all physical objects connected to the network. The IoT includes all things that have a unique id and a presence in an internet-like structure. Things may have an internet connection or tagged using Radio-Frequency Identification (RFID), Near Field Communication (NFC), etc.
- The *Internet of Locations* (IoL): refers to all data that have a spatial dimension. With the uptake of mobile devices (e.g., smartphones) more and more events have geospatial attributes.

Note that the IoC, the IoP, the IoT, and the IoL are partially overlapping. For example, a place name on a webpage or the location from which a tweet was sent. See also Foursquare as a mixture of the IoP and the IoL. Content, people, things, and locations together form the IoE as shown in Figure 1.

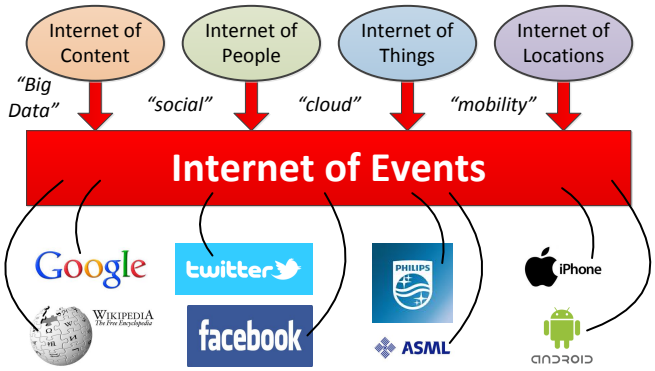

Fig. 1. The Internet of Events (IoE) is based on the Internet of Content (IoC), the Internet of People (IoP), the Internet of Things (IoT), and the Internet of Locations (IoL).

Data science aims to use the different data sources described in Figure 1 to answer questions grouped into the following four categories:

- Reporting: *What happened?*
- Diagnosis: *Why did it happen?*
- Prediction: *What will happen?*
- Recommendation: *What is the best that can happen?*

The above questions are highly generic and can be applied in very different domains. Wikipedia states that "Data science incorporates varying elements and builds on techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products" [21]. Many alternative definitions of data science have been suggested. For a short overview of the history of data science, we refer to [17].

The remainder is organized as follows. In Section 2 we discuss the unprecedented growth of (event) data and put it in a historical perspective. Section 3 compares data with oil, followed by Section 4 which discusses the value of this new oil. Section 5 describes the required capabilities of the data scientist. Section 6 lists some of the core technologies available to transform data into results. Finally, Section 7 describes the recently established Data Science Center Eindhoven (DSC/e).

2 Our Growing Capabilities to Store, Process and Exchange Data

Figure 1 describes the different sources of data contributing to the Internet of Events (IoE). As an example, take a modern smartphone like the iPhone 5S. As illustrated by Figure 2 such phones have many sensors. These may be used to collect data on a variety of topics ranging from location (based on GPS) to usage.

Fig. 2. Modern smartphones have many sensors that can be used to collect data.

It is difficult to estimate the growth of data accurately. Some people claim that *humanity created 5 exabytes (i.e., 5 billion gigabytes) of data from the Stone*

Age until 2003, that in 2011 that amount of data was created every 2 days, and that now (2013) it takes about 10 minutes to generate 5 exabytes [18]. The expanding capabilities of information systems and other systems that depend on computing, are well characterized by Moore's law. Gordon Moore, the co-founder of Intel, predicted in 1965 that the number of components in integrated circuits would double every year. During the last fifty years the growth has indeed been exponential, albeit at a slightly slower pace. For example, as shown in Figure 3, the number of transistors on integrated circuits has been doubling every two years. Disk capacity, performance of computers per unit cost, the number of pixels per dollar, etc. have been growing at a similar pace.

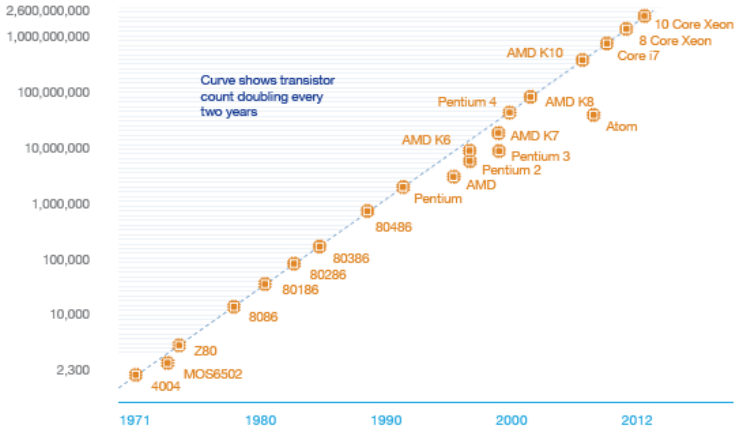

Fig. 3. Moore's law applies not only to the exponential growth of transistors on a chip: it also applies to processor speeds, communication speeds, storage space on hard disks, and pixels on a screen.

Note that Figure 3 uses a logarithmic scale: the number of transistors on a chip increased by a factor $2^{40/2} = 1048576$ over a 40-year period. To truly grasp this development, let us illustrate this using a few comparisons. If trains would have developed like computer chips, we could now travel by train from Eindhoven to Amsterdam in approximately 5 milliseconds (1.5 hours divided by $2^{40/2}$). Airplanes could fly from Amsterdam to New York in 24 milliseconds (7 hours divided by $2^{40/2}$), and we could drive around the world using only 38 milliliters of petrol. These examples illustrate the spectacular developments associated to Moore's law.

3 Big Data as the New Oil

Data science aims to answer questions such as "What happened?", "Why did it happen?", "What will happen?", and "What is the best that can happen?".

To do this, a variety of analysis techniques have been developed. However, such techniques can only be applied if the right input data is available. *Fancy analytics without suitable data are like sports-cars without petrol.* In fact, already in 2006 Clive Humby (co-founder of Dunnhumby) declared: *"Data is the new oil"*. However, only recently it became evident that data indeed represents incredible economic and societal value.

Using the metaphor "data=oil" we can definitely see similarities:

- *Exploration*: just like we need to find oil, we need to locate relevant data before we can extract it.
- *Extraction*: after locating the data, we need to extract it.
- *Transform*: clean, filter, and aggregate data.
- *Storage*: the data needs to be stored and this may be challenging if it is huge.
- *Transport*: getting the data to the right person, organization or software tool.
- *Usage*: while driving a car one consumes oil. Similarly, providing analysis results requires data.

So the different stages from exploring crude oil to using it to drive a car also apply to data science. However, there are also important differences between data and oil:

- Copying data is relatively easy and cheap. *It is impossible to simply copy a product like oil.* (Otherwise gas prices would not be so high.)
- Data is *specific*, i.e., it relates to a specific event, object, and/or period. Different data elements are *not exchangeable*. When going to a petrol station, this is very different; drops of oil are not preallocated to a specific car on a specific day. Production to stock of data is seldom possible. Typically, data elements are unique; therefore it is difficult to produce them in advance.
- Typically, *data storage and transport are cheap* (unless the data is really "Big Data"). In a communication network data may travel (almost) at the speed of light and storage costs are much lower than the storage costs of oil.

As pointed out before, Moore's law does not apply to classical means of transport by car, trans, or plane (cf. speed, fuel consumption, etc.). The end of Moore's law has been wrongly predicted several times. However, it is clear that the ultimate limits of the law come in sight. At some stage transistors cannot be made any smaller and clock speeds cannot be further increased. Therefore, the only way to keep up with the growing demands for storage and communication is to increase the number of computing entities. See the increasing numbers of cores in processors and the trend to use large clusters of commodity hardware in the context of Hadoop. Consider for example Google. Instead of relying on expensive proprietary hardware to store and process data, Google uses industry-standard servers that both store and process the data, and can scale without limits by using distributed parallel processing. Such massive parallelization results in a huge energy consumption. This is the reason why Google invests in renewable energy and decides on the location of its data centers based on the availability of energy sources.

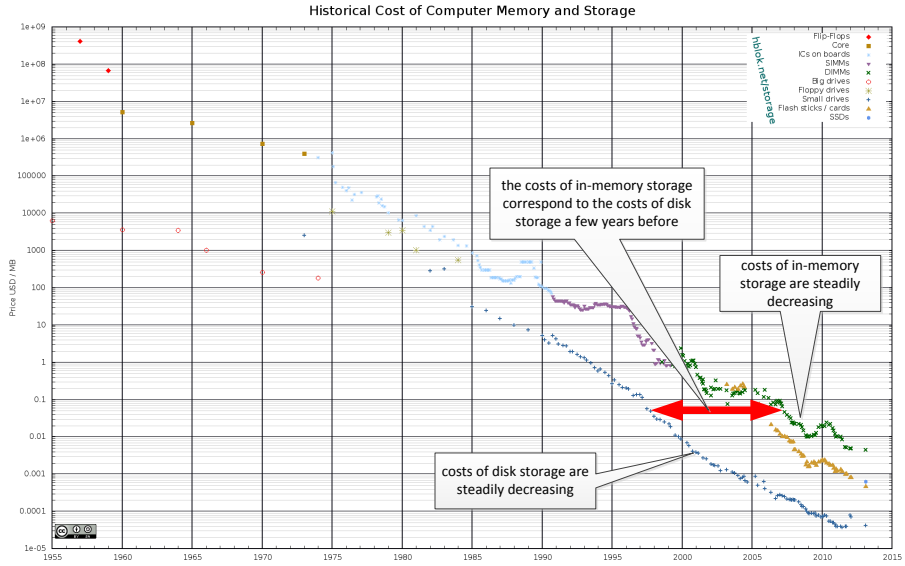

Fig. 4. Comparing the costs of different types of storage over time (taken from [13]).

Energy costs and the costs of hardware are also influencing the infrastructure most suitable for large-scale data science applications. Figure 4 shows the evolution of the costs of storage. The lower line refers to the decreasing costs of disk storage. However, as shown in Figure 4, the costs of in-memory storage are decreasing at a similar pace. Hence, the current prices of in-memory storage are comparable to the prices of disk storage of a few years ago. This explains the growing interest in in-memory databases and in-memory analytics. It now becomes affordable to load entire databases in main memory. The SAP HANA in-memory computing platform [16] is an illustration of this trend.

To understand the importance of storing data at the right place, consider the characteristics of the Xeon Intel chip shown in Figure 5. If the CPU requires a data element and it is available in it's L1 cache, then this takes only 1.5 nanoseconds. Assume that this corresponds to a distance of 90 centimeters. If the data is not in the L1 cache, but in main memory, then this takes 60 nanoseconds. This corresponds to a distance of 36 meters (using our earlier assumption that 90 centimeters equals 1.5 nanoseconds). If the data is not in main memory, but on a Solid-State Drive (SSD) then this takes 200.000 nanoseconds. This corresponds to a distance of 120 kilometers. To get the data from a regular hard disk takes 10.000.000 nanoseconds and corresponds to a distance of 6000 kilometers. Hence, shifting data from hard disk to main memory may result in incredible speed-ups.

Having the right "oil infrastructure" is crucial for data science. Moreover, innovations in hardware and software infrastructures (e.g., Hadoop) allow for types of analysis previously intractable. When using MapReduce techniques and distributed computing infrastructures like Hadoop, we are trying to optimize the

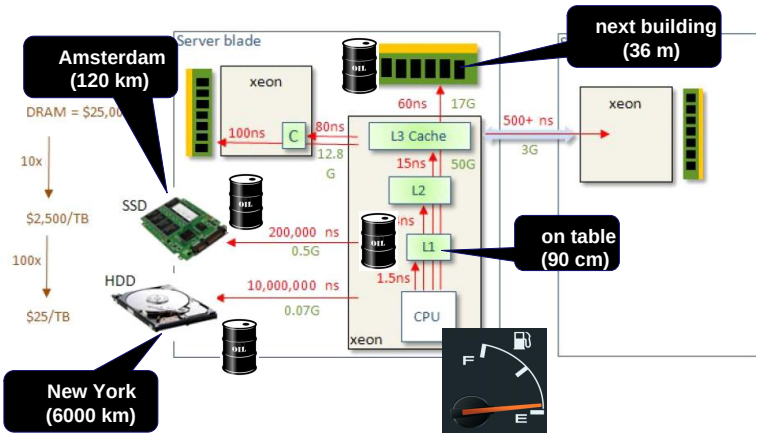

Fig. 5. How to get new oil? The power of in-memory computing becomes obvious by relating travel distances to the time required to fetch data in a computer.

alignment between data and computation (e.g., bringing computation to data rather than bringing data to computation).

4 On The Value of Data

In [4] the value per user was computed by dividing the market capitalization by the number of users for all main internet companies (Google, Facebook, Twitter, etc.). This study (conducted in 2012) illustrates the potential value of data. Most user accounts have a value of more than $100 dollar. Via the website `www.twalue.com` one can even compute the value of a particular twitter account, e.g., the author's twitter account (@*wvdaalst*) was estimated to have a value of $321. Adding up the different social media accounts of a typical teenager may yield a value of over $1000. Such numbers should not be taken very serious, but they nicely illustrate that one should not underestimate the value of data. Often the phrase "If you're not paying for the product, you are the product!" is used to make internet users aware of the value of information. Organizations like Google, Facebook, and Twitter are spending enormous amounts of money on maintaining an infrastructure. Yet, end-users are not directly paying for it. Instead they are providing content and are subjected to advertisements. This means that other organizations are paying for the costs of maintaining the infrastructure in exchange for end-user data.

The internet is enabling new business models relying on data science. Some examples:

– PatientsLikeMe.com connects patients having similar medical problems and sells this information to professionals. The community platform is based on the sharing of information that is resold to a third party.

– Groupon.com provides a broker platform where customers can get a discount by buying as a group. If the deal takes place, Groupon gets parts of the revenue.
– AirBnb.com connects people so that they can rent out spare rooms to one another. AirBnb gets commission.

In all cases data is used to connect people and organizations so that information, products, or services can be exchanged.

Besides enabling new business models, data science can be used to do things more efficient of faster. Moreover, data science plays a pivotal role in Customer Relationship Management (CRM). For example, data originating from different information sources (websites, sales, support, after sales, and social media) can be used to map and analyze the so-called *customer journey*. Organizations may use analytics to maximize the opportunities that come from every interaction customers have with them. Loyal customers are more cost effective to retain than acquiring new ones, since they are likely to purchase more products and services, are less likely to leave, and may help to promote the brand.

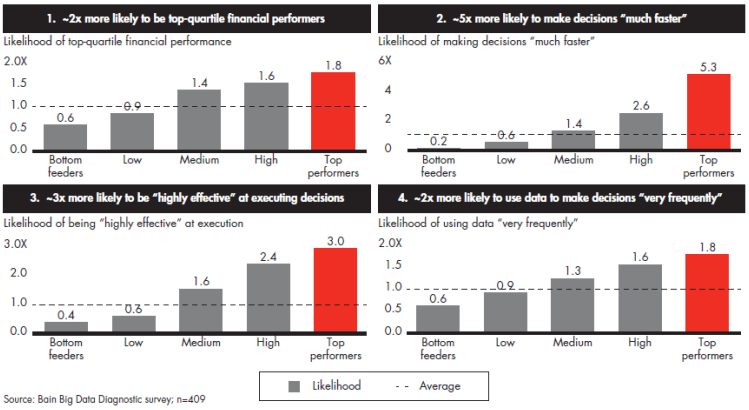

Fig. 6. Survival of the fittest: results of a Bain & Company study suggesting that companies with the best data science capabilities outperform the competition [15].

Optimizing the customer journey is one of the many ways in which organizations benefit from data science and extract value from data. Increased competition makes data science a key differentiator. Organizations that do not use data intelligently, will not survive. This is illustrated by various studies. See for example the results of a Bain & Company study [15] shown in Figure 6. We believe that in the future organizations will compete on analytics.

5 Data Scientist: The Sexiest Job of the 21st Century

Hal Varian, the chief economist at Google said in 2009: "The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?". Later the article "Data Scientist: The Sexiest Job of the 21st Century" [7] triggered a discussion on the emerging need for data scientists. This was picked up by several media and when analyzing job vacancies, one can indeed see the rapidly growing demand for data scientists (see Figure 7).

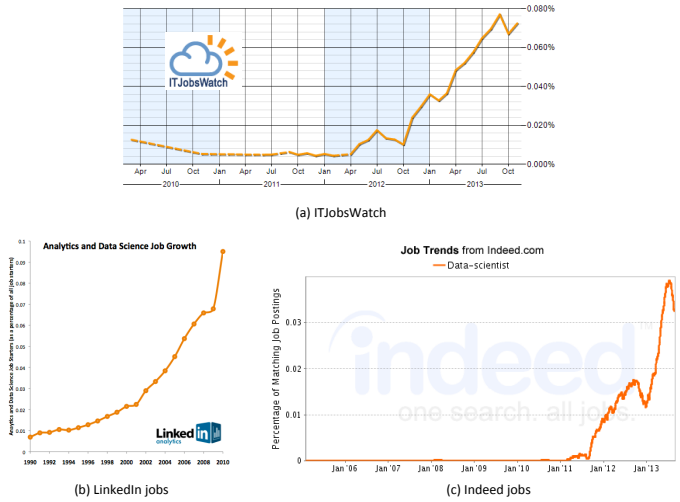

Fig. 7. The demand for data scientists is growing.

So, what is a data scientist? Many definitions have been suggested. For example, [7] states "Data scientists are the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information". Figure 8 describes the ideal profile of a data scientist. As shown, data science is multidisciplinary. Moreover, Figure 8 clearly shows that data science is more than analytics/statistics. It also involves behavioral/social sciences (e.g., for ethics and understanding human behavior), industrial engineering (e.g., to value data and know about new business models), and visualization. Just like Big Data is more than MapReduce, data science is more than mining. Besides having theoretical knowledge of analysis methods, the data scientist should be creative and able to realize solutions using IT. Moreover, the data scientist should have domain knowledge and able to convey the message well.

It is important to realize that data science is indeed a new *discipline*. Just like computer science emerged from mathematics when computers became abundantly available in the 1980-ties, we can now see that today's data tsunami is creating the need for data scientists. Figure 9 shows that data science is emerging

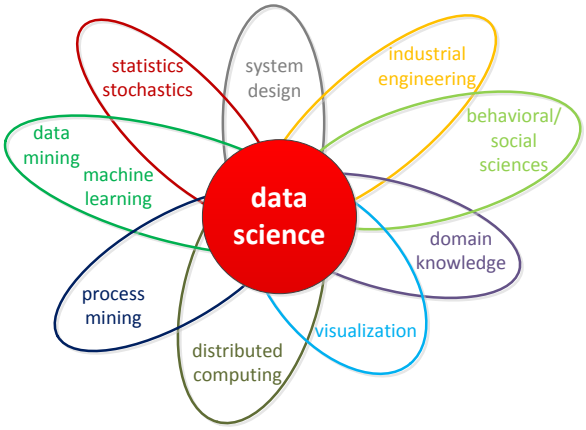

Fig. 8. Profile of the data scientist: different subdisciplines are combined to render an engineer that has quantitative and technical skills, is creative and communicative, and is able to realize end-to-end solutions.

from several more traditional disciplines like mathematics and computer science.

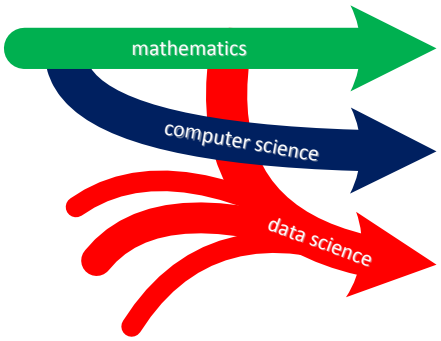

Fig. 9. Just like computer science emerged as a discipline when computers became widely available, data science is emerging as organizations are struggling to make sense of torrents of data.

6 Turning Data into Value: From Mining to Visualization

Although data science is much broader (cf. Figure 8) we would now like to briefly describe three "data science ingredients": data mining, process mining, and visualization.

In [8] *data mining* is defined as "the analysis of (often large) data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner". The input data is typically given as a table and the output may be rules, clusters, tree structures, graphs, equations, patterns, etc. Initially, the term "data mining" had a negative connotation especially among statisticians. Terms like "data snooping", "fishing", and "data dredging" refer to ad-hoc techniques to extract conclusions from data without a sound statistical basis. However, over time the data mining discipline has become mature as characterized by solid scientific methods and many practical applications [2, 5, 8, 14, 22]. Typical data mining tasks are *classification* (e.g., constructing a decision tree), *clustering*, *regression*, *summarization*, and *association rule learning*. All of these are based on simple tabular data where the rows correspond to instances and the columns correspond to variables.

Process mining aims to *discover, monitor and improve real processes by extracting knowledge from event logs* readily available in today's information systems [1]. Starting point for process mining is an *event log*. Each event in such a log refers to an *activity* (i.e., a well-defined step in some process) and is related to a particular *case* (i.e., a *process instance*). The events belonging to a case are *ordered* and can be seen as one "run" of the process. Event logs may store additional information about events. In fact, whenever possible, process mining techniques use extra information such as the *resource* (i.e., person or device) executing or initiating the activity, the *timestamp* of the event, or *data elements* recorded with the event (e.g., the size of an order).

Event logs can be used to conduct three types of process mining [1]. The first type of process mining is *discovery*. A discovery technique takes an event log and produces a model without using any a-priori information. Process discovery is the most prominent process mining technique. For many organizations it is surprising to see that existing techniques are indeed able to discover real processes merely based on example behaviors stored in event logs. The second type of process mining is *conformance*. Here, an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. The third type of process mining is *enhancement*. Here, the idea is to extend or improve an existing process model thereby using information about the actual process recorded in some event log. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model. For instance, by using timestamps in the event log one can extend the model to show bottlenecks, service levels, and throughput times.

Data and process mining techniques can be used to extract knowledge from data. However, if there are many "unknown unknowns" (things we do not know

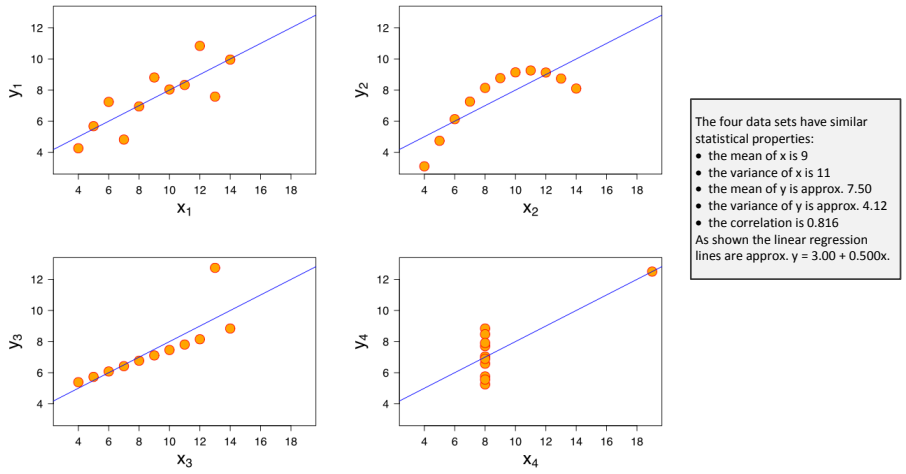

Fig. 10. Anscombe's Quartet [3]: Although the four data sets are similar in terms of mean, variance, and correlation, a basic visualization shows that the data sets have very different characteristics.

we don't know), analysis heavily relies on human judgment and direct interaction with the data. *Visualizations* may reveal patterns that would otherwise remain unnoticed. A classical example is Anscombe's Quartet [3] shown in Figure 10. The four data sets have nearly identical statistical properties (e.g., mean, variance, and correlation), yet the differences are striking when looking at the simple visualizations in Figure 10.

The perception capabilities of the human cognitive system can be exploited by using the right visualizations [20]. Information visualization amplifies human cognitive capabilities in six basic ways: 1) by increasing cognitive resources, such as by using a visual resource to expand human working memory, 2) by reducing search, such as by representing a large amount of data in a small space, 3) by enhancing the recognition of patterns, such as when information is organized in space by its time relationships, 4) by supporting the easy perceptual inference of relationships that are otherwise more difficult to induce, 5) by perceptual monitoring of a large number of potential events, and 6) by providing a manipulable medium that, unlike static diagrams, enables the exploration of a space of parameter values [6, 19].

The term *visual analytics* was coined by Jim Thomas to advocate a tight integration between automatic techniques and visualization. Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [11]. For example, data and process mining can be used in conjunction with interactive visualization.

7 Data Science Center Eindhoven (DSC/e)

In 2013, the *Data Science Center Eindhoven* (DSC/e) was established as Eindhoven University of Technology's (TU/e) response to the growing volume and importance of data. About 20 research groups of the Department of Mathematics & Computer Science, the Department of Electrical Engineering, the Department of Industrial Engineering & Innovation Sciences, and the Department of Industrial design of TU/e are involved in this center.

In line with the TU/e policy, DSC/e's research contributes to the challenges of the TU/e Thematic Research Areas: Health, Energy, and Smart Mobility. Each of these areas witnesses a rapid growing volume of data triggering a variety of scientific challenges. Data science is also highly relevant for the high-tech industry in the Brainport region ("the smartest region in the world"). However, DSC/e is not limited to the TU/e's thematic research areas or the Brainport region. In fact, industries such as the financial industry and the creative industry heavily depend on data science.

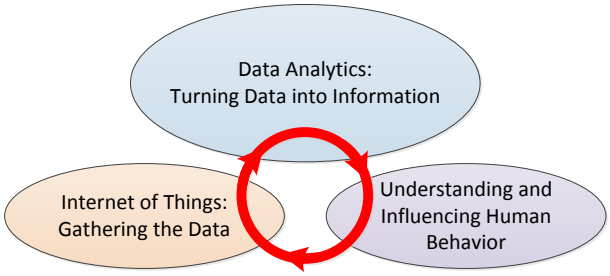

Fig. 11. The three main research lines of DSC/e.

TU/e has strong research groups in areas related to data science: computer science, mathematics, electrical engineering, industrial engineering, innovation sciences, and industrial design. In subdisciplines such as process mining, which are at the very heart of data science, TU/e is globally leading. The DSC/e aims to further strengthen research in three broader areas (Fig. 11):

- *Internet of Things: Gathering the Data*
- *Data Analytics: Turning Data into Information*
- *Understanding and Influencing Human Behavior*

DSC/e's research focuses on developing new insights (models, theories, tools) to be able to add and extract value from real sets of heterogeneous data. On the one hand, the groups involved will continue to conduct focused research in particular areas relevant for data science. On the other hand, the DSC/e initiative will fuel multidisciplinary research combining expertise in the different DSC/e research groups contributing to DSC/e.

Given the empirical nature of data science, DSC/e collaborates with a wide range of organizations. Collaborations include larger joint research projects, PhD projects, master projects, and contract research. Examples of organizations collaborating within DSC/e are Philips, Adversitement, Perceptive Software, Magnaview, Synerscope, and Fluxicon.

References

1. W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin, 2011.
2. E. Alpaydin. *Introduction to Machine Learning*. MIT press, Cambridge, MA, 2010.
3. F.J. Anscombe. Graphs in Statistical Analysis. *American Statistician*, 27(1):17–21, 1973.
4. B. Bergstein and M. Orcutt. Is Facebook Worth It? Estimates of the Historical Value of a User Put the IPO hype in Perspective. MIT Technology Review, `http://www.technologyreview.com/graphiti/427964/is-facebook-worth-it/`, 2012.
5. M. Bramer. *Principles of Data Mining*. Springer-Verlag, Berlin, 2007.
6. S.K. Card, J.D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, 1999.
7. T.H. Davenport and D.J. Patil. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, pages 70–76, October 2012.
8. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT press, Cambridge, MA, 2001.
9. M. Hilbert and P. Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025):60–65, 2011.
10. C. Howard, D.C. Plummer, Y. Genovese, J. Mann, D.A. Willis, and D.M. Smith. The Nexus of Forces: Social, Mobile, Cloud and Information. http://www.gartner.com, 2012.
11. D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age: Solving Problems with Visual Analytics*. VisMaster, http://www.vismaster.eu/book/, 2010.
12. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2011.
13. J.C. McCallum. Historical Costs of Memory and Storage. `http://hblok.net/blog/storage/`.
14. T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
15. T. Pearson and R. Wegener. Big Data: The Organizational Challenge. Bain and Company, San Francisco, `http://www.bain.com/publications/articles/big_data_the_organizational_challenge.aspx/`, 2013.
16. H. Plattner and A. Zeier. *In-Memory Data Management: Technology and Applications*. Springer-Verlag, Berlin, 2012.
17. G. Press. A Very Short History of Data Science. Forbes Technology, `http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/`, 2013.
18. R. Smolan and J. Erwitt. *The Human Face of Big Data*. Against All Odds Productions, 2012.
19. J.J. Thomas and K.A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE CS Press, 2005.

20. J.J. van Wijk. The Value of Visualization. In *Visualization 2005*, pages 79–86. IEEE CS Press, 2005.
21. Wikipedia. Data Science. `http://en.wikipedia.org/wiki/data_science`, 2013.
22. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.