The Future of Data

Peter Apers

Link to Video recording

0. Preliminary remarks

[video 09m 37s] Welcome to all of you. Welcome to the members of the Executive Board. Welcome to my colleagues from within and outside the university. Welcome to my friends and family. I would like to welcome all of you into my living room. If you squeeze in a bit, everyone will fit.

I announced my departure via LinkedIn. I have never received this many "likes" in my life. At first, I thought it was great. Later on, I realised perhaps it meant that people wanted me gone.

We do not really know the meaning of those likes, yet we assign a meaning to them anyway. That is what Data Science is all about. Today, I want to discuss the development of Data Science with you, which is of course closely related to the development of the internet and the emerging role of Data.

1. Introduction

[video 10m 33s] Let me begin with the internet, which has become fully integrated into our daily lives. At first, it was a fun way to send emails and share pictures. That was all quite innocent.

- Now, we frequently hear about fake news and the influencing of our opinions in ways that we did not expect. We are used to commercials designed to influence our purchasing behaviour. However, we could not have predicted the rise of fake Facebook accounts created to propagate political ideas and influence elections;
- Now, we hear about dynamic pricing: prices change between the morning and the afternoon and buying online via an Apple device is more expensive than using a Windows device. Are we heading towards personalised pricing, where the price you pay depends on what retailers know about you: where you work, what car you drive, where you live? Insurance companies are already experimenting with this: your registered behaviour on the road affects your car insurance premium.

- Now, we hear about apps that help us detect serious illnesses such as skin cancer ourselves at an early stage. Another example is coughing into your smartphone and it telling you if you have pneumonia.

One thing is clear: the internet has rocked our society to its core. Especially when we know that fake news is given preference on social media. We see the world of data becoming more professional and business-like. How do we respond to that development?

Today, I want to talk to you about how the world of data was able to develop so rapidly and what the current trends are. I will use a few timelines for this:

- Developments in Data;
- Developments in IT;
- My own activities.

2. Data is timeless

[video 13m 07s] Ladies and gentlemen, I started my talk by discussing the internet in our daily lives. Now, I want to move on to my own area of expertise: data. Data is timeless. It is always about recording and structuring data and then making decisions based on the answers to questions about the data.

Let's start with recording. During my inaugural speech in 1986, I already talked about quipus. Quipus were used by the Incas to keep inventory of e.g. corn and beans. As you can see, a quipu consists of knotted coloured strings. It was long believed that quipus were only used to record numerical information. However, after discovering two long-lost quipus, the anthropologist Sabine Hyland proposed that quipus are actually narrative letters. The number of colour and knot combinations used in quipus is practically the same as the number of symbols of a logographic script, similar to the Egyptian hieroglyphics and the Chinese characters. We have not yet managed to decipher the quipus, because we lack a Rosetta Stone that would allow for translation into a known language.

After recording comes structuring. Structuring data is about recording the relationships between data. Take a mother-child relationship, for instance. It is an example of a hierarchical relationship or a 1:N relationship. A mother can have multiple children, yet a child can only ever have one biological mother. I am starting to sound like Miss Ank from *De Luizenmoeder*.

Of course, a text document is also data in which we recognise a certain structure. As far back as 1945, Vannevar Bush did something remarkable. He wanted to include references in one document to parts of other documents. To do so, he used Memex and "As we may think." The concept of hypertext was born. Today, we associate this idea with the World Wide Web, our internet. That did not exist yet, though, which made this a truly revolutionary idea.

I already mentioned that the goal of registering and structuring data is to make decisions, e.g. about the redistribution of corn supplies. That is only possible if you can ask questions about the data. For this, you need a description of the data (which we call a schematic) and a query language.

3. Late 1960s (1969-1970): the beginning of data storage and communication

[video 16m 12s] I have now brought you into my study. For the first databases, the data model, in which the schematic was described, resembled the data structures that we use in our programming languages. That would soon change.

The relational data model was introduced in the late 1960s. The essence of this model was that relationships were no longer recorded in data structures, but rather by matching the attribute values of tuples. As you can see in the image, this was a great idea. However, it resulted in several processing methods at the storage structure level to answer questions. One method might take a lot longer than another. The goal is to process the question in as little time as possible. The area of expertise I worked in within the field of databases concerned the efficient processing of questions.

Around the same time as the introduction of the relational data model, the first computers were linked together via the ARPANET. These computers were located at UC Los Angeles and Santa Barbara, Stanford University and University of Utah. Many more would follow later. Next, the Internet Protocol was introduced to allow computers to communicate with each other. This marks the beginning of the internet as we know it today, where data is distributed across millions of computers. Remember that none of this existed just fifty years ago.

The late sixties marked the beginning of the storage and communication of data.

4. First IT systems (1974-1990): the start of digitisation

[video 18m 03s] Midway through the 1970s, the first IT systems were introduced. Ingres, the first relational database system, was developed at the University of California in Berkeley by Mike Stonebraker. At roughly the same time, IBM was working on its System R, another relational database system. Large-scale IR systems were also being built around this time, such as Roger Summit's Lockheed Dialog System.

I started my doctoral research at the VU Amsterdam in the late seventies. The subject was the optimisation of query processing and data storage in distributed databases. Storing data on multiple interconnected computers was still a futuristic idea at the time.

In the late 1980s, Dutch database researchers collaborated with Philips on the development of a large parallel database machine: the PRISMA machine. It had one hundred nodes and kept all data in its main memory. It made a huge international impact. Around that time, Annita Wilschut wrote the most acclaimed article in my group. It was about the efficient implementation of the hash join in a pipeline architecture. Although I cannot delve too deeply into the matter, take it from me that her work is found in most database systems to this day.

The internet was still in its infancy: the first computers were being linked together, it became possible to send email messages (just don't ask me how) and the first index programs became available. I can still remember taking a kind of sabbatical at the CWI at the end of my first deanery. Like an amazed child, I was exploring the structure of the internet. Back then, it was quite difficult to find what you were looking for, as Google did not exist yet.

I already mentioned the birth of hypertext in 1945. It became a reality during this time. Suddenly, it was possible to include a reference in one document to a different document stored on a computer halfway around the world. The complex processes involved, e.g. how to find other computers and how computers communicate with each other, were hidden in the background. This contributed enormously to the popularity of the internet with the masses.

Database researchers faced a difficult time. The popularity of the internet created quite the identity crisis in the database world. The internet was all about data, yet this data was not being stored in a database. The database schema was thrown out the window, as was the closed world assumption and proper query languages. Gone were all the goodies from

the database world. As the internet became increasingly popular, the interest in and appreciation of IR, the basis for the search engines we know today, grew.

Let me illustrate this identity crisis. I was invited to attend a workshop for top researchers in Como. As you can imagine, I was honoured. We would be talking about complicated topics: databases and the internet. It was daunting to come up with something truly challenging to present to such an esteemed company. I conferred with my AIOs. During this time, you could send files from your PC to the printer in the hallway. We came up with the idea of storing data on the internet, instead of on your PC.

Mike Stonebraker, the database pioneer I mentioned before, was relentless in his criticism: "complete nonsense, no company will ever store its data on the internet." Many other attendees supported him. During the break, as I went to the toilet, a few sympathetic colleagues told me that some ideas are presented too early. The concept of "data in the cloud" did not exist yet. Today, it is hard to imagine a world without cloud storage.

Looking back, it is clear that the initial step towards digitisation was taken during this time of the first IT systems.

5. The new millennium (1990-2010)

[video 22m 54s] Ladies and gentlemen, perhaps you still remember the first internet bubble that burst in the year 2000. It might have cost you a lot of money at the time. Yet there was something else happening around the dawn of the new millennium, remember? The infamous Y2K bug. In an effort to save memory, only the last two digits of a year were saved. For 1999, only 99 was saved. For 2000, only 00. The problem is that 00 comes before 99, while 2000 comes after 1999. Everyone predicted total chaos. Luckily, quite a few IT companies made good money off the crisis and life as we knew it continued.

Remember when Booking.com launched in 1996? Google was introduced in 1997 and Bol.com in 1999. In less than twenty years, these organisations have become major factors in our lives, each in their own way. This trend continued unabated after 2000; Wikipedia in was launched in 2001, Facebook in 2004 and the first iPhone in 2007.

When you sum it up like this, you can clearly see the enormous impact that IT technology has had and continues to have on our society. No other form of technology has impacted so many people so quickly.

If a reporter asked people on the street in 1998 if they wanted a mobile phone, they looked at him in surprise. Why would they want to be available all the time? Today, we experience serious withdrawal symptoms if we ever forget our smartphone.

I view this wave of innovation as the first IT tsunami. We are currently on the eve of a second tsunami: that of Big Data. The first wave was a technological one, while the second will be more about content. It is hard to predict what the third wave will be. I expect that the collection of data will lead to breakthroughs in many fields and that we will enter a phase of "personalised solutions," e.g. in the healthcare sector. IT companies will play a prominent role in this sector.

I was asked to give the dies speech in 2003. We shot a few videos to illustrate what IT would bring us in the future. You can see a summary of one of these videos here. Note that everything we predicted back then, only fifteen years ago, has now been fully integrated into our daily lives.

It was high time for the world of computer science and IT researchers to start organising itself better, e.g. to earn recognition as a scientific discipline. In 1997, the first NOAG-I (Nationale OnderzoeksAgenda Informatica, National Research Agenda Computer Science) was hosted by the SION board, which I was the chairman of at the time. In 2005, it was followed by the NOAG-ICT. This was the first time that the field united behind a limited number of larger themes. The concept of the two-stage rocket was born then. During the first stage, these themes were added to the scientific research agenda. During the second stage, researchers could submit research projects pertaining to these IT themes. The pie got bigger, so everyone could have a piece.

Let's look ahead for a moment. The following initiatives also contributed substantially to the realisation of a higher degree of organisation in the IT field:

- 3TU.NIRICT
- ICTRegie
- COMMIT
- EIT Digital, known as ICT Labs back then,
- ICT Topteam
- COMMIT2Data

Nearly everyone who was or is involved in these projects is here today. Thank you very much for your contributions.

As a result of the major IT-related changes in society, the government began to realise that IT research had to be funded in the Netherlands as well. Just think of the many FES and Bsik projects, as well as the many projects that were part of the EU's framework programmes. **Now, in 2018, after all ICT FES projects have been completed and the funding has run out, the demand for research into digitisation is once again growing. There is hope yet. All that remains is to find the money.**

6. The development of Data Science

[video 29m 15s] I would like to discuss the rise of Data Science with you. When I started out as a researcher in 1976, I never expected the quantity of data to explode like it did. We realise now that data contains something very important: knowledge. This explosion of data will rock our society and the worlds of business and science in the years to come.

Jim Gray was a leading scientist in the Database field. He laid the foundation for data-intensive science. He talked about the fourth paradigm of science. After the first phase of experimentation came the second phase of forming theories. It was followed by the third phase of simulations and now data-intensive science, in which large quantities of data are processed using tools in order to gain new insights and reach new breakthroughs. Just think of radio astronomy.

I briefly got to work with him when I was the Editor-in-Chief of the VLDB Journal. Jim worked for Microsoft R&D at the time, which had a one-man location in San Francisco. That is where Jim lived and where his sailboat was docked. He received the Turing Award for his work.

I talk about Jim in the past tense. He never returned home from a trip to spread his mother's ashes off the coast of San Francisco. Just a few days after this tragic incident, many people were studying satellite footage of the Pacific Ocean off the coast of San Francisco in an effort to discover what happened to him. Despite everything, this is a good example of data-intensive research. Looking through hundred of thousands of satellite pictures to find a forty-foot sailing boat is like looking for a needle in a lot of haystacks. That is the challenge of Data Science: finding knowledge in a mountain of data.

To overcome this challenge, Machine Learning techniques are used.
This video shows a simple neural network during the learning phase. One example is teaching a car to brake when it detects certain things. At the bottom, you see the three detection spheres, all represented by either a zero or a one. The output sphere is at the top and the desired output is seen to the left (brake or not), also represented by a zero or a

one. The spheres in the middle row are constantly modified to approach the desired output. That is called learning. As you can see, the chance of a wrong decision decreases as the learning process goes on. To simplify the matter, the knowledge is in the middle spheres of the neural network at the end of the learning process. The car can now decide on its own whether to brake or not. This is an extremely simple example of using Machine Learning to teach a car about road safety.

The number of applications for extracting knowledge from data and then applying this knowledge is enormous. As a result, there is a growing demand for Data Scientists. Emile Aarts showed that this presents a fine challenge for all universities. Willem Jonker also plays an important role. With EIT Digital, he not only takes on this challenge at the European level, but also includes the business development side.

Let's all work together to create top-quality Data Science study programmes to meet the enormous demand for qualified data scientists. We should do so before our data is abused for purely commercial purposes. **There are countless other challenges that require further study. To name but a few: ownership, transparency, ethics, legality and scalability.**

7. Trends in Digitisation

[video 34m 30s] Now it is time to look towards the future. The data explosion is the driving force behind the digitisation of our society. This, in turn, is made possible by the ongoing decentralisation of the internet. This decentralisation turns us all into producers of data. The digitisation of our society will only pay off if our society's system innovates along with it. I would now like to talk to you about these three trends: decentralisation, producers and system innovation.

7.1 Decentralisation of IT

Let's begin with the trend of decentralisation. In 1943, Thomas Watson said: "I think there is a world market for maybe five computers." In the end, his estimate was slightly off. When I was twelve or thirteen, my father took me to KLM's computer centre to look at the new IBM 360 mainframe. Gerrit Blaauw, emeritus professor of this university, was a co-designer of the 360 series.

After the large mainframes came the PCs. There was also a time of barely portable PCs. The first iPhone was launched in 2007. Smartphones are now followed by wearables. At the moment, the market's main focus is the sports sector. However, the time when

wearables are integrated into your clothing is not far off. The fashion world cannot wait to open up this new market.

Your smartphone and wearables are collecting all kinds of data for apps: the number of steps you take, the number of floors you walk up, your GPS location, whether you are standing up or sitting down, your heartrate, whether you are exercising, running, cycling, driving, etcetera. Your smartphone is basically a very powerful sensor. You might still think it is a phone, although it is quite old-fashioned to use it as such. It is millions of times more powerful than the IBM 360 and very much smaller (like comparing a car to something held in the palm of your hand). **Conclusion of the decentralisation trend: nearly all computational power is now found in the capillaries of the internet, from which data is retrieved constantly. This is the perfect example of decentralisation.**

This is a good time to reflect on what electrical engineering has achieved in terms of miniaturisation in just fifty years. If we were to build an iPhone 5S using the transistors from the sixties, it would be the size of the Eiffel tower and require two nuclear power plants to operate (thanks to Bram Nauta for this apt comparison). Remember that the transistor was invented in 1947. That was just one transistor in a lab. Today, you are carrying around hundreds of millions or even billions of transistors in your smartphone, your wearables, etcetera. We rarely stop to think about this innovation and expect the trend to continue for years to come. It is important to realise that this innovation is the result of an enormous academic and industrial effort. **Decentralisation was only possible because of this miniaturisation.**

7.2 Consumers become producers

The second trend is that of consumers becoming – consciously or unconsciously – producers of data with their smartphones and wearables.

The question is how we should handle this development. There are two perspectives: either we believe that the data is ours and we should be free to decide what to share or we share everything and let scientists and businesses do what they need to do to further improve our quality of life.

The first perspective requires systems to protect our data. The second perspective is about value creation. This is a fundamental choice that is further complicated by the ethical and economic interests involved. For example, how do we handle ownership? Does the data collected by a wearable belong to the manufacturer or the owner of the device?

A growing number of IT companies is collecting medical data. Perhaps these IT companies know when something is wrong with your health, even before you or your GP do. Google can already predict with a great degree of accuracy whether a patient will die in the next forty-eight hours. How very interesting! Do they have a reporting obligation for these insights? Do we even want to know? In the future, more and more medical breakthroughs will come from IT companies.

Now that we realise that our smartphones and wearables mainly collect data, the question is whether we should even have to pay for a device that collects our valuable data. Perhaps our data creates more value in the business model of a health app than what it costs to manufacture the device. In other words, where will Apple's profits come from in five years' time: from the sale of IT products or the collected data? To take the question one step further: will Apple still be an IT company in five years' time, or will it become a health company with an IT division?

Conclusion of the data production trend: how do we control the data streams and what do we get in return as individuals and as a society?

7.3 System innovation

The third trend has to do with system innovation. The decentralisation in the IT world is also found in e.g. the energy sector. A shift is taking place from large power plants to solar panels on your roof, i.e. from centralised to decentralised energy generation. Of course, this is part of the effort to realise a transition from fossil to sustainable energy resources. As with data, this makes producers out of consumers.

Let me tell you a story. There are many small-scale energy cooperatives. One evening, while sitting in a bar, the director of the local pool says that he will be forced to close the swimming pool as he can no longer pay his energy bills. A farmer sitting next to him says that he is not being paid for the wind energy he generates on his land. By working together, these two can solve each other's problems. Of course, this is a made-up story, yet it does raise some questions: should energy be returned to the network without limitation, for what fee and who decides on this fee? It demonstrates that decentralised energy generation calls for an update of our laws and regulations.

In the healthcare sector, as in the energy sector, IT innovations result in improvements that do not always fit well within our slightly outdated and closed-off system of laws and regulations. Taking full advantage of digitisation often requires a system innovation. It will be much easier for countries without any micro-scale regulations to benefit from

digitisation. The ultimate question is therefore where we want the Netherlands to stand in five years' time?

Conclusion of the system innovation trend: let politicians show leadership by facilitating system innovations and creating room for the implementation of IT innovations that improve our quality of life.

Looking at all three trends, it is clear that decentralisation has made the collection of data possible. Business models are updated based on the insights gained from the collected data. Oftentimes, the bottleneck is not caused by a lack of creativity, but rather by the limitations of our system of laws and regulations that changes at a much slower rate. In all three areas, steps must be taken to improve our quality of life.

8. Considerations

[video 45m 12s] I have cleared out my study to make room for my successor Joost Kok. Let's return to my living room.

I want to leave you with something to ponder. When you look at the developments of the internet from a distance, you can see the following developments taking place:

1. First, there was a technological development that we did not really know what to do with. The first files were exchanged and the first emails were tentatively sent;
2. Next, search features were introduced and the first products were sold online;
3. Then came the data platforms (Google, social media, etcetera);
4. We are now in the phase of collecting data from individuals;
5. The next phase will involve the development of new products and services based on the insights gained from the collected data.

We are on the eve of the Big Data tsunami. All ingredients are there. What will we do with it? Who will benefit from it? I expect it will be the people who understand that data contains knowledge and who update their business model accordingly in a country whose government facilitates this transition.

Here is something to think about over drinks.

Consideration 1: collecting data

At the foundation of collecting data is an entire chain that begins with sensors, sensing and sensor networks, linked via the Internet of Things. The applications are legion:

robotics, autonomous vehicles, smart industry, personalised healthcare, sustainability. The University of Twente has a strong presence throughout the entire chain, as a result of a collaboration between nanotechnology, IT, business administration and behavioural science and a wide range of applications. You should seize the opportunity to develop a partnership with the business world.

Consideration 2: sharing data

Our society is ruled by data platforms from the United States. These data platforms are used to collect and share data. These days, Facebook knows more about the citizens of e.g. Enschede than its city officials do. It is high time that we invest in European data platforms, if only to learn more about our own people and safeguard the quality of our society. There is no point to building a European version of Facebook, but it is important to develop specific and secure platforms for the healthcare sector. If you don't do it, Amazon or some other company will.

Consideration 3: updating business models

In France, mailmen and -women check on senior citizens to see if they are doing okay as they make their rounds – for a fee, of course. That demonstrates a clear understanding of the fact that a postal company's main asset is that they visit every citizen on a daily basis. That is another form of value creation. Looking at the logistical data from a new perspective creates a win-win situation for businesses and our society.

Let me invite you to come up with at least one new service for your own organisation during the reception. For inspiration, here is something to think about: will Apple be more than an IT company in five years' time? Will the NS be more than a railway company? Will a university be more than an educational institute?

In the Netherlands, we continue to underestimate the impact that digitisation will have on our society. When it comes to matters of IT, we demonstrate a conservative reflex. Now is the time to take action. Otherwise, we risk being the only ones who fail to understand what the colourful quipus, which I mentioned at the start of my talk, can tell us.

To illustrate, during my final meeting with the computer science chairs, Boudewijn Haverkort quoted from my inaugural speech. He concluded that the call to action I voiced at the time is still relevant today.

I was glad to hear that I was not spouting nonsense at the time, yet disappointed when I realised how little progress we have made in all that time. Sometimes, it looks like nothing more than a drop in the ocean.

9. Thanks

[video 49m 54s] I want to conclude by thanking you.
First of all, I want to thank anyone who inspired and assisted me during the preparation of this speech. I also want to thank the speakers of the symposium. They made it perfectly clear that Data Science still presents us with some fine challenges.

Secondly, I want to thank the university with its ecosystem for the inspirational environment it offered me for thirty-three years. A never-ending stream of new students, new AIOs, new employees, new projects, new applications, new rectors and new structures. It was always challenging and inspirational and I was never bored for a moment.

Thirdly, I want to thank my national and international colleagues, both those in my own field and those working in related fields. I am grateful for the recognition that I received as a scientist, together with my PhD students, and I appreciate the facilitating role I got to play for scientists in other fields, both within and outside the university.

Fourthly, I want to thank Arnold Smeulders. It is great to have a friend who does not have his own agenda and dedicates himself entirely to the IT community. It was a pleasure to work with you and you taught me a lot. I also want to thank Amandus Lundqvist. We never would have made it this far without him.

Finally, I want to thank Berta and the kids for trying to make me a better man, even though I hardly ever had the time.

I have spoken.

Video progress:
52m:22s Thom Palstra speaks and briefly introduces Arnold Smeulders.
53m:23s Arnold Smeulders addresses Peter.
01u:05m:20s Thom Palstra addresses Peter.
01u:07m:32s Thom gives flowers to Berta Apers.

01u:07m:54s End of the ceremony, followed by several announcements and the cortège's departure.