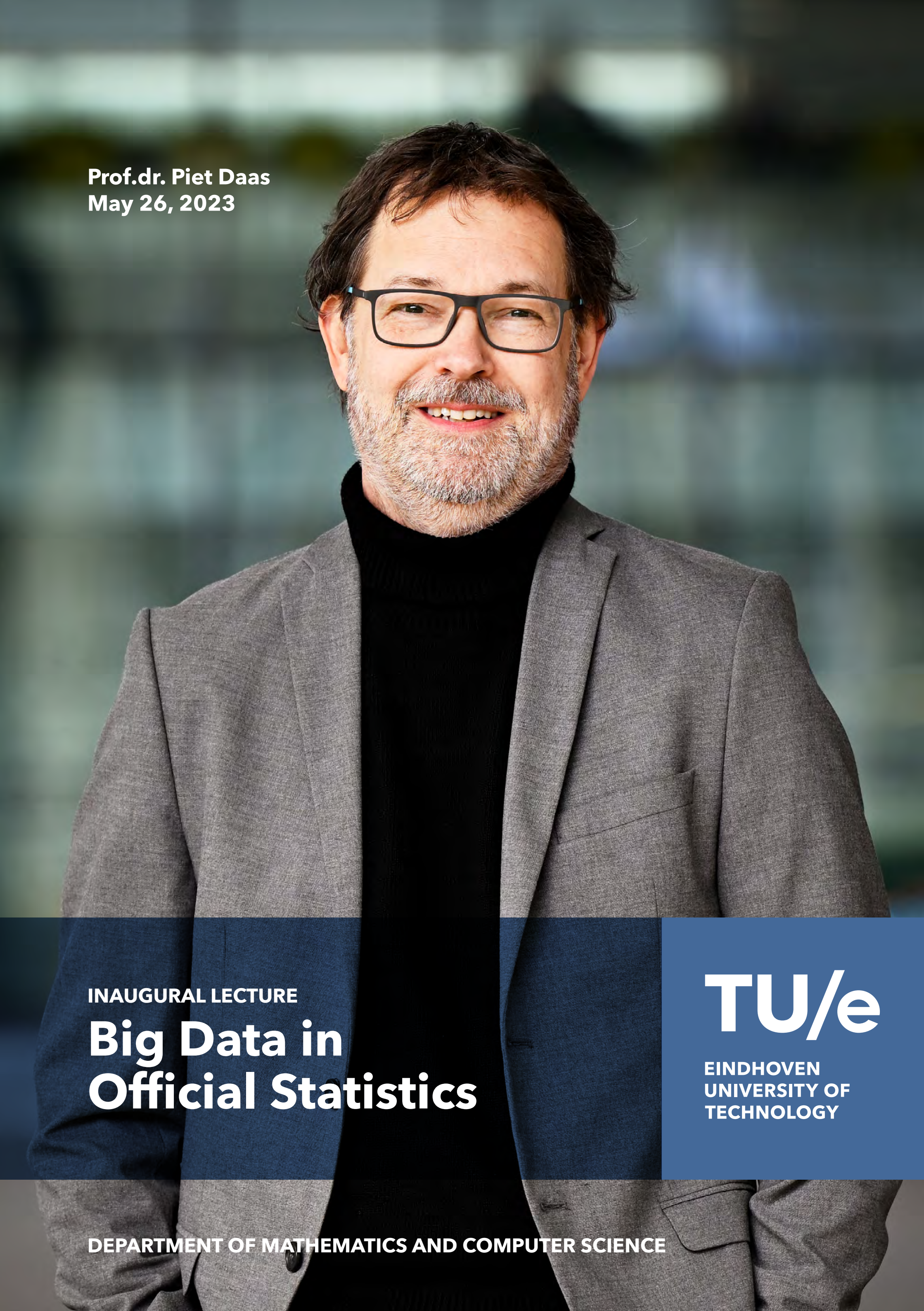
Big data in official statistics



Prof.dr. Piet Daas

May 26, 2023

INAUGURAL LECTURE

Big Data in Official Statistics

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

PROF.DR. PIET DAAS

Big Data in Official Statistics

Presented on May 26, 2023
at Eindhoven University of Technology

Introduction[1]

Dear family, friends, and colleagues,

Welcome and thank you for coming to Eindhoven to listen to my inaugural lecture. An inauguration is an introduction of a new professor to the public and the university. In this lecture, I will explain the purpose of my chair, which has been installed by Statistics Netherlands and Eindhoven University of Technology. It is positioned within the Statistics group in the Department of Mathematics and Computer Science. The topic of my chair is ‘Big Data in Official Statistics’, which contains terms that may seem vague to you at the moment. I will start by introducing both terms and discussing the opportunities and challenges in the area of Big Data - for official statistics - and illustrate them with examples.

OFFICIAL STATISTICS

Let’s start by introducing what official statistics are. When you watch the news on TV or other devices or read newspaper articles, you are often confronted with statistics. Some of these statistics are created by commercial agencies, some are made up, and some are of high quality. Official statistics are statistics that are, almost by definition, produced by governmental agencies or other public bodies, such as Statistics Netherlands, the Dutch statistical office. Official statistics provide an indispensable element in the information system of a democratic society by serving it with data about important themes, such as the economic, demographic, social, and environmental situation of a country. Because of this, it should be obvious that official statistics must adhere to a number of fundamental principles; a total of 10 have been defined to assure impartiality, reliability, accountability, and transparency [1]. Hence, the data and methods used to extract information from data and produce statistics should be of the highest standards possible and should follow scientific principles and professional ethics. You can’t just produce a number and simply assume that this act alone makes it correct.

[1] I thank Marco Puts and Yvonne Gootzen for their valuable feedback on earlier drafts.

The two data sources that are currently predominantly used in official statistics production are survey data [2], which is usually obtained from a representative sample of people, companies, or products (let's call them units), and administrative data [3], which is data that has (already) been collected by another organization. Sample surveys are under the control of the statistical office, which does its utmost to use data from a representative subset of the target population, i.e., a subset that reflects the characteristics of the target population as accurately as possible. Administrative data are collected and maintained by other (usually governmental) organizations, such as data collected by the Dutch tax office. They usually contain a large subset of the target population and sometimes even, by definition, the complete target population. Statistics Netherlands has been using sample surveys ever since the first experiments were performed in 1924 (in Brabant, by the way) [4]. In the Netherlands, the use of administrative data for official statistics started in 1981 for the Virtual Census [5]; in Dutch: 'de virtuele volkstelling'. After that, administrative data were used in increasing amounts in the production of other official statistics. It is, however, important to realize that before statistical institutes used sample surveys and administrative data as sources of input, they all applied a census-oriented way of working. This meant that everyone and everything of interest in the country was (attempted) to be enumerated as completely as possible [4]. Counting took place of, for example, the number of men eligible to serve in the army, the number of cows, or the amount of grain harvested.

BIG DATA

So, what about Big Data? We need to realize that nowadays, in our modern world, enormous amounts of data are being generated by all kinds of electronic devices, systems, and people online. These data are not thrown away but usually remain in storage. It is these data that are commonly referred to as 'Big Data' and can (potentially) be used for a whole range of new (unforeseen) purposes [6], including official statistics [7]. But let's start by defining it. There are various definitions available online, which already provides a clue about the heterogeneity of the term and its use. I think that the definition of Big Data that best describes it in the context of official statistics is:

> *"Big data are usually (extremely) large datasets that can contain both structured and unstructured data and that, when analyzed computationally, may reveal patterns, trends, and associations relating to the behavior and interactions of the units included."*

With unstructured data, a reference is made here to data sources such as texts and images. These data sources are completely clear to us as humans. However, for a computer, these data sources have a less well-defined or even undefined structure and meaning. Big Data brings texts and images into the realm of official statistics. From the definition, it is also clear that the challenge for Big Data lies in obtaining patterns, trends, and associations in a reliable and reproducible way. Extracting and using information from Big Data is a relatively new area for official statistics. The first report of a National Statistical Institute on the potential of Big Data for official statistics was published in 2009. It was actually a report by me and my Statistics Netherlands colleagues Marko Roos and Marco Puts [8]. In that document, we talked about the potential of "new data sources" and it was the onset of the Big Data hype within the official statistical world. As such, it was a great report, although I still think that we should (also) have published an English version.

So, what are the overall benefits that Big Data offers compared to the data sources currently used in official statistics, i.e., survey and administrative data? The potential benefits are:

i) Speed: Big Data is continuously being generated, which means that statistics can be produced at a much higher frequency (e.g., weekly or daily) and with less delay.
ii) More detail: Big Data may contain data at a very high level of granularity which could enable the production of statistics at a level of detail not possible with traditional sources.
iii) Newness: Big Data may contain information on currently unobserved phenomena and, as such, enable the production of statistics on completely new topics.
iv) Burden reduction: Any information extracted from Big Data may not have to be collected by other means. This could seriously reduce the administrative burden on persons and organizations that currently provide that information.

These are all arguments to decide to thoroughly study the potential use of Big Data for official statistics production. And that is what has been done since 2009. As a result, a considerable number of applications have emerged and many experimental studies have been performed. Table 1 provides an overview of both the official and experimental statistics that predominantly make use of Big Data. These statistics are either produced by National Statistical Institutes (or comparable organizations) or are the result of attempts to produce nationally relevant numbers by others.

Table 1. Overview of Big Data-based statistics in production and of an experimental nature[a]

No.	Product	Big Data source(s) used	Off/Exp[b]	Countries[c]	Freq.[d]
1	Consumer Price Index (inflation index)	Scanner data, web-scraped prices (± surveys)	O	Multiple	M (2W)
2	Biodiversity trends (incl. butterfly index)	Internet observations (+ survey)	O	NL, EU	Y
3	Traffic intensity statistics	Road sensors	O	NL	M
4	Internet economy	Websites (+ admin data)	O	NL	Y
5	Online platform statistics	Websites (+ survey)	O	NL	Y
6	Land use, crop/vegetation detection	Satellite/aerial pictures (± admin data)	O/E	CA, AU/Multiple	Y (Q)
7	Public transport monitor	Public transport smart card (+ survey, admin)	E	NL	Once
8	Mobility patterns (during COVID)	Mobile network operating data	E	Multiple	D
9	Job vacancy/advertisement statistics	Online job ads	E	Multiple	M
10	Enterprise characteristics	Websites	E	Multiple	Y
11	Daytime population/commuting stat.	Mobile phone data, transport data	E	Multiple	M
12	Innovative tourism statistics	Multiple sources (e.g., websites, road sensors)	E	Multiple	M
13	Social media sentiment	Social media messages	E	Multiple	D,W,M
14	SDGs (incl. urbanization)	Satellite/aerial pictures	E	EU, UN	Once
15	Electricity/energy consumption	Smart meter data	E	ES, DK, NO, UK	D
16	Maritime and inland waterway stat.	Automatic identification system data	E	NL, GR, PL, UN	Once
17	Innovative company detection	Websites	E	NL, DE, BE	Once
18	Outbound tourism statistics	Mobile network operating data	E	ES, FI, AT	Once
19	Solar panel detection	Aerial pictures	E	NL, BE, DE	Once
20	Suicide numbers/mental health index	Weblogs, Twitter	E	KR, ID	D
21	Accommodation statistics	Data from most important booking platforms	E	EU	M,Y
22	World Heritage sites popularity	Wikipedia page views	E	EU	Once
23	Social mood on economy Index	Twitter messages	E	IT	D
24	Social unrest indicator	Social media messages	E	NL	D
25	Retail sales index	Debit card transaction data	E	NO	Once
26	Road accidents	OpenStreetMap (+ population, vehicle fleet)	E	IT	Once
27	Caravan home identification	Real estate websites (+ address register)	E	UK	Once
28	Travel flows	Oyster card data (+ census data)	E	UK	Once
29	Determining residence and mobility	Twitter data	E	UK	D
30	Effect of ships on underwater life	Automatic identification system (+ geo) data	E	NL	Once
31	Monthly postal trade statistics	Postal receptacle identifiers	E	UN	M
32	Monthly global trade statistics	Global trade data	E	UN	M
33	Trade volume nowcasts	Automatic identification system data	E	UN	D
34	Air transport index (during COVID)	International civil aviation organization data	E	UN	M
35	Migration indicator	Mobile phone data, social media data	E	UN	Q,Y
36	Displacement and disaster statistics	Mobile phone data	E	UN	Event, Y
37	Information society statistics	Mobile phone data, internet connection speed	E	UN	Y
38	Early economic indicator (spending)	Payment card transaction data	E	US	D
39	Economic sentiment index	Google Trends data	E	CH, IMF	D
40	Poverty/income/demography indicator	Google Street View images	E	Multiple	Once
41	Environmental statistics, water	Geospatial and earth observation data	E	UN	Once
42	Skills of graduates	Aggregated LinkedIn data (+ statistics)	E	NL	Once
43	Environmental health indicator	Emission and noise data (+ geo data)	E	NL	Once
44	Economic activity detection	Websites	E	AT, NL	Once

a To be included, the study must use Big Data as the major source or as a very important input source and must concern official or national statistics.

b Official (Off, O) and Experimental (Exp, E).

c The two-letter abbreviations used are country codes, Europe (EU), and United Nations (UN), IMF is International Monetary Fund, and Multiple is used to indicate that the topic has been studied in more than five countries.

d Frequency abbreviations used are Daily (D), Weekly (W), Monthly (M), Quarterly (Q), and Yearly (Y).

A more detailed version with links to the individual studies is available online (http://pietdaas.nl/table1). The author has tried to create as complete an overview as possible. Barteld Braaksma is gratefully acknowledged for reviewing the table and for additional suggestions. Please inform me if (you think that) a study is absent by sending me an email, including a URL to the online report/webpage of that study. The email address can be found on the tables webpage.

Table 1 shows a total of 44 Big Data-based statistics, of which six have officially been published. All others are of an experimental nature and, as indicated, have either been produced once or more frequently. The overview makes clear that the majority of the statistics listed are experimental, indicating that they are either not yet considered of sufficient quality to become 'official' or are merely a first try at something new. The table additionally illustrates the huge variety of Big Data sources used. Typical examples are webpages, road sensors, social media data, bank card transactions, satellite pictures, and mobile phone network operating data.

In my opinion, the reasons why the number of officially published Big Data-based statistics (in Table 1) is rather low are twofold. The first is related to the particular properties of Big Data, which require the use of methods unfamiliar to many official statisticians, such as using texts and images as input data. The second reason is a non-methodological one. It indicates the need for a paradigm shift [9] or at least an update [10] within the official statistical world. This 'shift' is not only related to the acceptance of the use of new methods or new ways of working, of which some are specific to Big Data, but also in the way that Big Data is looked upon in general. This is, however, not an unexpected observation as similar remarks were made when the use of sample surveys was first considered by official statistical offices in the 1900s. The Statistics Netherlands discussion paper on 'The Rise of Survey Sampling' by Professor Emeritus Jelke Bethlehem describes this clearly [4]. On page 13 of that paper, in which he gives an overview of the rise of survey sampling in general and for the Netherlands in particular, he states that:

> "*It took almost 30 years [before sampling] was approved as a valid statistical method. … The distrustful attitude of statisticians towards sampling was not surprising. Until the end of the 19th century, they emphasized the importance of complete enumeration at every opportunity*."

So, it took a while before producing statistics from (probability) samples became universally accepted within National Statistical Institutes. Currently, this is the default way of working (and thinking) at statistical offices, even though the use of administrative data is highly encouraged [11]. In that sense, we may still have 15 years to go before Big Data becomes a more acceptable and known source. More importantly, the statement makes clear that the way one looks at the data sources that one currently uses, which are survey and administrative data at this point in time, highly affects one's judgment when looking at new sources (and their opportunities), such as Big Data [12]. This is a very important point and we should also be aware that it takes considerable time to find out how one produces statistics with new data sources in a reliable and reproducible way. And, to top it all off, this needs to be done in a critical and transparent way.

From the above, it is clear that Big Data is a very diverse and broad topic to study [13]; this is one of the reasons I like it so much. Just look at the diversity of Big Data sources listed in Table 1. It contains data sources that include numbers, texts, images, and more. As I already mentioned, many Big Data sources are easy to interpret by humans (like images and texts) but not by computers. Methods to reliably extract information from these kinds of data sources are still under development [14-16]. A research area like artificial intelligence plays an important role in this. I also haven't even touched upon dealing with privacy in the context of Big Data or the IT environment needed to efficiently work with huge amounts of data. These two are certainly essential topics, but because of time constraints they will not be discussed in this lecture. For completeness, I have included two appendices providing an overview of these topics in the digital version of this lecture.

As mentioned before, I stated that the way of working with Big Data differs somewhat from the way that many statisticians are familiar with when working with data. This is not only the result of the usually extremely large amounts of data available but also the fact that the availability of these large amounts highly encourages a data-driven way of doing science. Looking for patterns or trends in data is an example of this. However, contrary to what one would expect, this is not the most common way that many (official) statisticians work. Note that this should not be interpreted as a negative remark. The point I want to get across here is that there are essentially two approaches to doing science; let it be clear that these are merely two sides of the same scientific 'coin'. The two approaches are referred to as inductive and deductive reasoning and they simply indicate different starting points for doing science. The first uses a data-driven way, making it very interesting for Big Data-based studies, while the latter uses a theory-driven way of working. Inductive reasoning aims to generalize observations into a theory; it moves from specific observations to broader generalizations. Deductive reasoning starts with a theory and, by formulating a hypothesis, tests the correctness of that theory on specific cases; it moves from more general to more specific. Both approaches interact at the theory and observation level and hence form a cycle [17]. Both are valid ways of working, but when large datasets are available, *starting* in an inductive way is an obvious choice. The other (big) advantage of data-driven work is that one may run into unexpected findings – things not (yet) included in a theory. In the end, the ultimate goal for anyone working with Big Data, from whichever viewpoint, should be the development of a 'Big Data theory' [18]. With such a theory, I mean a generalized way of using the information included in Big Data to enable the reliable (and reproducible) extraction and production of results, such as (official) statistics. This is obviously a long-term goal.

WORKING IN A BIG DATA WAY: AN EXAMPLE

Let's illustrate the data-driven way of working with an example. Here, we focus on answering the question: when was the term Big Data first used? There are a number of ways to obtain the information needed to answer that question, including a literature study [19], but we will start with the obvious choice: using an online search engine, such as Google [20]. When you type that question in the Google search bar, the following response is obtained.

Google search: *"Some argue that [Big Data] has been around since the early 1990s, crediting American computer scientist John R Mashey, considered the 'father of big data', for making it popular. Others believe it was a term coined in 2005 by Roger Mougalas and the O'Reilly Media group. …"*

For completeness, and because you can't give a presentation nowadays without at least mentioning ChatGPT (an artificially intelligent chatbot) [21], I've asked the same question to this oracle of wisdom:

ChatGPT (v3): *"The term 'Big Data' was first used in the early 2000s, although its exact origin is not clear. Some sources attribute the term to John Mashey, a computer scientist who used it in a presentation in 1997. Others attribute it to Doug Laney, an analyst at Gartner, who used the term in a report in 2001. …"*

From these somewhat different answers, it seems obvious that the term 'Big Data' originated somewhere between the early 1990s and the mid-2000s, but both answers indicate that its exact origin is unclear. But now comes the interesting part: how would one answer this question in the best Big Data-driven way possible? For me, it would be obvious to start looking at documents (papers or books) written at various points in time and count how often the term 'Big Data' occurs in all of the documents published each year. You could do this by going to the library, ordering all books available according to their publishing year, and start counting the occurrences of the term. But this takes a tremendous amount of time and effort. Fortunately, there is an easier (online) option available. Interestingly, Google has scanned many, many books over the years, extracted the text from these scans (with a technique called Optical Character Recognition), and made the extracted texts available for searching online. The texts of all books scanned can be searched via the Google Books Ngram Viewer website [22] (what's in a name?). Let's see what happens when we search for the occurrence of the term 'Big Data' in English books from 1880 to 2019 (for completeness: case-insensitive and smoothing set to zero). This result is shown in the insert of Figure 1.

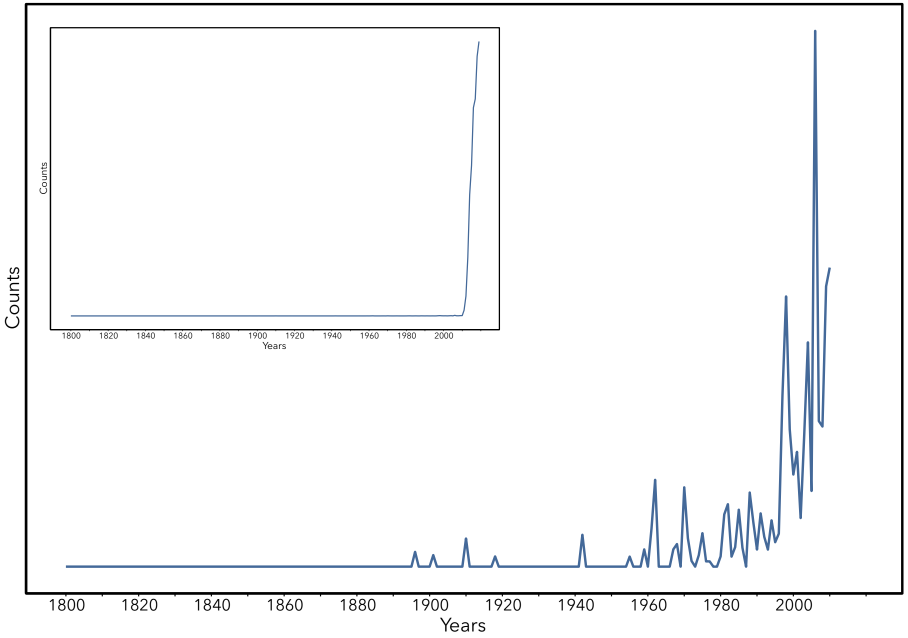


Figure 1. Results of searching for the term 'Big Data' in Google Books Ngram Viewer from 1800 to 2009 (main) and from 1800 to 2019 (insert).

From the insert in Figure 1, it is obvious that the term 'Big Data' was increasingly used from 2010 onwards. But since we are predominantly interested in the period before that, let's limit the search to the year 2009. The results of this search are shown in the main plot of Figure 1. What this figure suggests is that the term 'Big Data' was certainly used before 1990 and became increasingly popular over the years. I won't discuss the individual peaks (and books) here, but it is important to realize that the results indicate occurrences of the term 'Big Data' as early as the 1900s. A detailed checking of the texts extracted from the individual books included reveals that some of these peaks are not actually the result of 'Big Data' and are incorrectly assigned. This is caused by i) Optical Character Recognition errors, which means that an error occurred when the text was extracted, resulting in the erroneous creation of the term ‘Big Data’; such as from the words ‘birthday’ (handwritten) and ‘big dam’, or ii) the fact that a fairly recently published book was wrongly assigned to an earlier year. The first actual occurrence of the term ‘Big Data’ was in a book published in 1912, where it was used as a synonym for large amounts of data. This continued until the 1980s. From then on, it was predominantly used to indicate the challenge of analyzing large amounts of data on computers with limited memory, up until the 1990s. After that, the term ‘Big Data’ started being used in the context of the definition I gave earlier on. In that sense, the answers of Google and ChatGPT are not entirely wrong; they are, however, obviously incomplete and miss the starting point.

The reasons for discussing this example are threefold. The first is simply to illustrate the data-driven way of working. The second is to show that Big Data can provide better (and valid) answers to questions when the results are checked carefully and that this needs to be done according to scientific principles. The third is to also make clear that Big Data can (and does) indeed contain errors [23], which need to be dealt with accordingly. This is a bridge to the next topic: the quality of Big Data.

Quality and Big Data

The quality of the input data used and the quality of the output derived from it are essential to (official) statistics production. In the case of Big Data, this becomes an even bigger challenge as many of these sources are ‘generated’ by non-governmental organizations, usually active in the private sector. Hence, not a lot is often known about the underlying process from which the data originate. This seriously affects its use. Numerous papers have been written on Big Data and quality and a considerable number of quality frameworks have been developed for its use. I refer to the paper by me and my colleagues Yvonne Gootzen and Arnout van Delden [24] as an example of this, which also provides an overview. Such frameworks usually include many quality aspects and I won’t discuss them all in detail here. I will focus on the three most important ones (in my opinion). These are quality aspects related to i) the concept measured, ii) the population included, and iii) stability over time. Obtaining information on each of these key aspects is essential to determining if a Big Data source can be used for (official) statistics production.

i. CONCEPT

With the concept, I mean the phenomenon that one intends to measure, such as the price of a product or the age of a person. These are both examples of directly measurable concepts, which can - for instance - be obtained from a webpage of a company or from someone’s social media profile page (either directly or derived from their birthdate). However, there is also a way to indirectly obtain a concept. I’ll illustrate this with a biological example. For instance, when a researcher wants to count the number of (emperor) penguin colonies in the Antarctic, they could obtain that information by visiting the continent and counting the number of penguin colonies observed. This takes considerable effort for this continent as one needs to visit all potential locations for a complete overview. There is, however, an alternative Big Data way to do this. One could use high-resolution satellite pictures of the Antarctic, which are freely available online, and look for any traces that those colonies leave behind [25]. Figure 2 shows an example of such a picture.
In the picture, clear darker-colored areas on a white (snowy/icy) background can be observed.

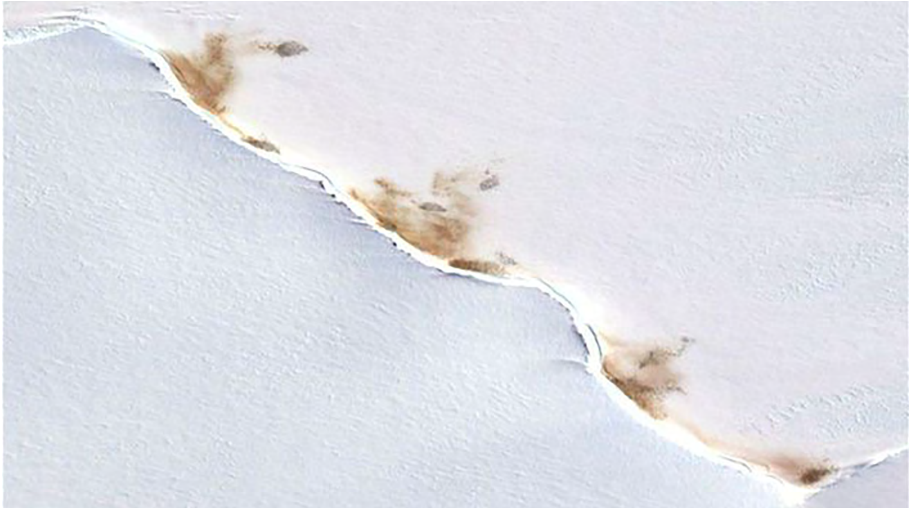


Figure 2. Satellite picture of an Antarctic region displaying emperor penguins (dark brown) and their excrement (light brown) - image by QuickBird satellite (© 2018 Digiitalglobe, inc.)

The dark brown regions are penguins, but the 'smudgy' light brown areas observed are the guano stains of these birds, i.e., 'penguin poop'. The light brown patches provide an indirect way to detect the presence of penguin colonies. In fact, researchers have demonstrated that by using this means of (indirect) measurement, more penguin colonies (61) are found compared to the traditional way of measuring (54). Such an approach is also possible for concepts of interest to official statistics. Examples of this include detecting online platforms from website texts [26] and deriving socioeconomic characteristics from Google Street View pictures [27]. This indirect use seriously increases the potential applications of Big Data but also introduces some potential new causes of error. In the context of the penguin colony example, the light brown patches could, for instance, be caused by the 'poo' of another species of penguin or the result of something completely different, such as algae growth or environmental pollution. In both cases, the number of colonies will be overestimated. Here, validation studies are needed. These are studies that specifically check if the Big Data-based findings have been correctly interpreted. They can also be used to detect if the indirect relation observed remains stable over time (see point iii).

ii. POPULATION

The population for which data are included in a Big Data source is the next important quality aspect. This isn't only about coverage, i.e., the part of the target population included in the source, but also about *how well the population of units included represents the target population* (for the intended statistic). When the target population is not completely included, it is essential to correct in some way or another - for these compositional differences to assure that the Big Data-based findings are not biased. Traditionally, this is done by comparing the background characteristics of the units in the source with those of the target population. However, for many of the Big Data sources typically studied with official statistics (Table 1), such characteristics are either absent or only very limitedly available. If possible, it is preferred to include the data of as many units as possible (and deduplicate) [26, 28]; this is a census-oriented approach. There is a need for alternative approaches here, such as combining Big Data with other types of sources (see also the section on combining sources) or the other way around [29], applying time series-based methods [30], applying correction methods used in clinical trials [31], using subpopulation-based groups [32], or developing workable non-probability-based correction methods [33, 34]. Eurostat, the statistical office of the European Union, has produced a report with an overview of a whole range of methods that could potentially be used [35].

iii. STABILITY OVER TIME

The stability of the source and the results obtained over time is the third quality aspect I'm discussing. This is essentially about the stability of the Big Data-based findings over time. One should be aware that Big Data sources are subject to change (often caused by a change in the behavior of the people/units generating the data), which can affect the findings derived from these sources. Let's illustrate this with an example: the detection of innovative companies based on website texts [28]. This is a study that plans to measure the concept of innovation in an indirect way. Based on the data from the Community Innovation Survey, a Statistics Netherlands survey on a sample of 10,000 businesses, a set of innovative and non-innovative companies was selected and the texts on their websites were obtained and studied. Without going into too much detail, this work revealed that it was possible to identify innovative companies based on the occurrence of particular combinations of words on their website with an accuracy close to 90%.

Very interesting and detailed results were obtained, including an almost exact replication of the official estimate of the number of large innovative companies (the target population of the survey) and an estimate of the number of small innovative companies (which were not included in the survey) [28]. Hence, the findings were not only validated but also provided new insights. As an important side note, similar studies have been performed in other countries, which confirmed that website texts could also be used to detect innovation in countries like the UK and Germany [36] with accuracies of around 70%. The reason I'm telling you this is that upon checking the findings of our Dutch study at later points in time, it was observed that the association found between the website's text and innovative companies slowly deteriorated over time [37]. This is illustrated in Figure 3.

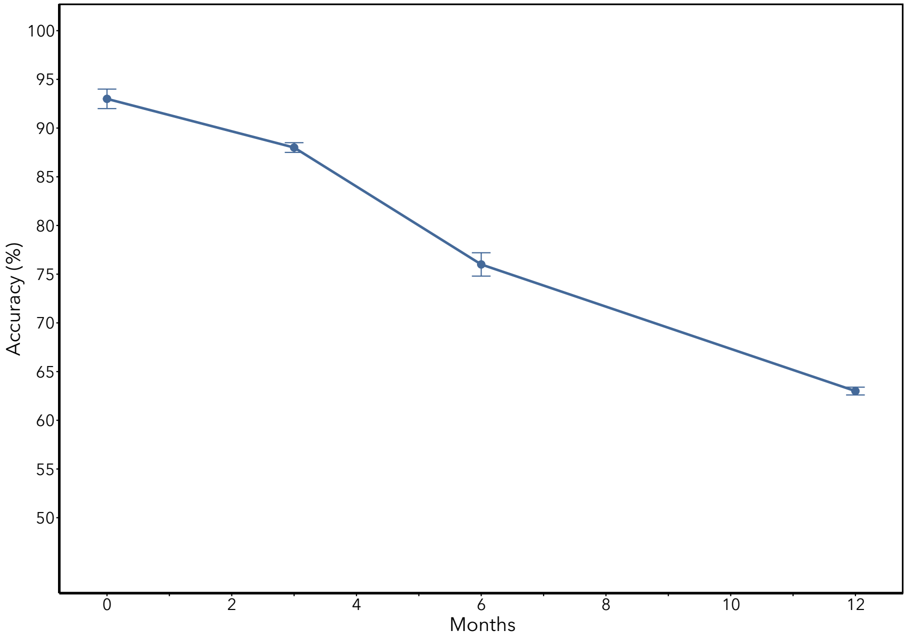


Figure 3. Accuracy and standard deviation of the innovation detection model, developed at month 0, for the websites of companies used for model development collected at various points in time. The results of 100 independent tries are shown.

The phenomenon shown in Figure 3 is commonly referred to as 'concept drift', although 'model degradation' is a better way to describe it [37]. In this particular case, it is extremely challenging to solve. Simple retraining of the model[2] does not suffice, so it's not only the association between the words and the concept of innovation that has 'drifted'. We recently found that the original classification of (some of) the companies used to train the model also changed over time. Learning how to deal with such changes is essential to assuring that particular Big Data sources can (remain to) be used for official statistics in the long run. A similar phenomenon occurred in the famous Google Flu Trends study [38, 39]. In this example, a subset of the words included in Google search queries was used to derive the onset and spread of flu in the population of a whole range of countries (from 2008 to 2015).

[2] Here, 'model' refers to a set of mathematical equations that establishes relationships between variable values in the data. Such a model enables one to, for instance, calculate the probability of a website belonging to an online platform business based on the words in the text [26]. The term 'algorithm' can be used as an alternative.

Combining sources

When one looks at the successful applications listed in Table 1, it becomes apparent that many of them are the result of combining Big Data with other sources. There are a number of reasons why this works really well but one of them is certainly the fact that this enables one to deal with any compositional difference between Big Data and the target population. Combining sources with Big Data is therefore an important topic to study, and this is what my colleague Yvonne Gootzen is investigating in her PhD research. She is studying, in a systematic and logical way, how the metadata (data about the data) of various sources can be used to combine Big Data with surveys and/or administrative data. The first result is a quality framework that should be applied during the design phase of a (new) statistic [24]; I have discussed three of the essential points included just a few minutes ago. After applying the framework and selecting which sources can be used, an exploratory analysis of the data sources is recommended to confirm the assessments of the framework and select which methods are most suitable. The ultimate goal of this work is the development of a (software) tool that can do all of this automatically and provide the user with possible (path)ways in which the sources could be combined. We expect that this fundamental work will greatly stimulate the use of Big Data (for official statistics) and improve the quality of the statistics derived from it.

Let's illustrate this with a case study, one on the topic of mobility in which we focus on people traveling by car. In this study, a total of (in principle) four data sources are combined. Two of them are Big Data sources, viz. road sensor and OpenStreetMap data. The other two sources are survey data from the Dutch National Travel Survey (ODiN) and (a combined set of) administrative data. The latter contains background characteristics of persons, such as the locations where people live and work [40]. Figure 4 provides a schematic overview of the sources, the important variables used in combining them, and the order in which they are combined. The figure is included to illustrate the complexity of this topic.

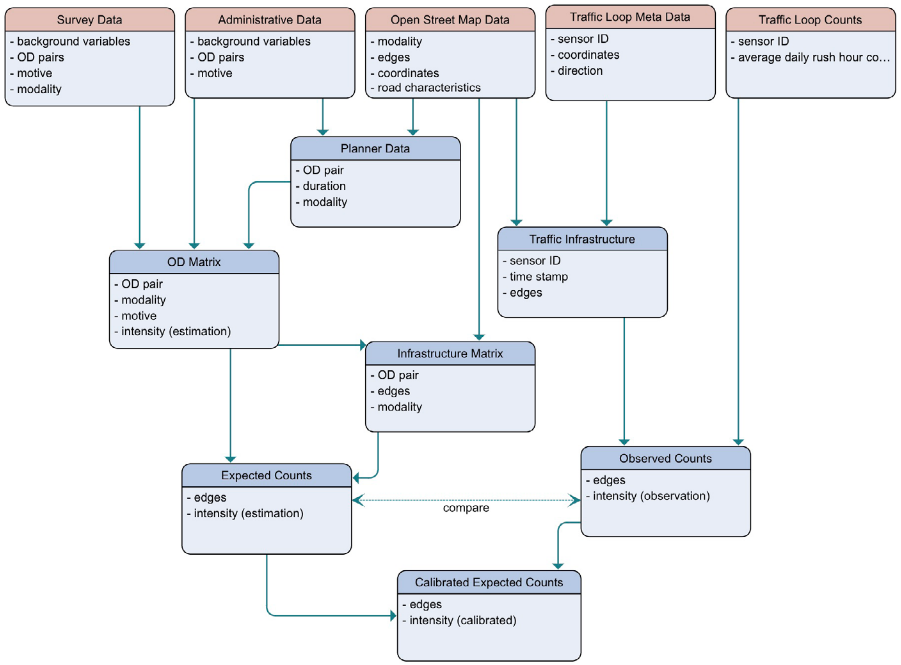


Figure 4. Overview of the ways in which two Big Data sources, a sample survey, and an administrative dataset are combined in the mobility study [40].

The intended mobility statistic, people traveling by car, is the result of the following stepwise approach. First, the survey data was used to determine the probability of a certain transport modality given the background characteristics of a person. Obviously, for this study, only the modality 'car' is of interest. Secondly, those probabilities were applied to the combined set of administrative data (containing the background characteristics of persons and their start and end locations), which was subsequently aggregated into an origin-destination (OD) matrix. The OD matrix is composed of pairs of neighborhoods and the expected number of people that travel to work by car. Third, OpenStreetMap data was used, in combination with OpenTripPlanner software, to convert the OD pairs into routes consisting of road segments. This resulted in an expected traffic intensity of cars for each road segment. Essentially, the route planner acted as a converter between the two unit types (neighborhoods and road segments). Finally, the available minute-based road sensor data were filtered based on vehicle length to make sure that only cars were included (and longer vehicles such as buses and trucks were excluded). The data was then aggregated to one intensity value for each sensor by taking the sum of all observations during the morning rush hour peak. By doing this, most travel from home to work was taken into account while minimizing the inclusion of travel for leisure or travel from work to home, as these predominantly tend to occur outside of the morning rush hour. The intermediate result of these steps is a dataset including two variables for all road segments that have road sensors: expected intensity and observed intensity. Assuming the observed intensity to be the ground truth, a model was subsequently trained that calibrates the expected intensity for a sensor to be closer to the observed value of that sensor. The model is applied to all of the road segments (including the ones with no road sensors and, as a consequence, no observed intensities) to produce a calibrated value of the expected intensity [40]. These results are shown in Figure 5. This example makes clear that new and much more detailed statistics can be produced by including Big Data and combining it with more traditional data.

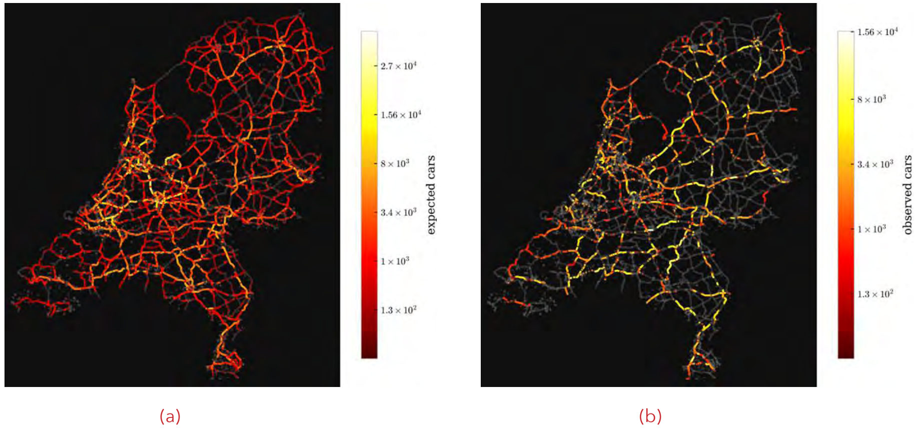


Figure 5. Expected (a) versus observed (b) traffic counts on the Dutch road network (from [40]). Note that the legends do not use the same value range. Since not all road segments contain sensors, such as local roads, (a) contains much more detailed results compared to (b).

Data Science methods

Working with Big Data, and working in a data-driven way in general, has been greatly stimulated by the recent onset of the academic field of Data Science. This relatively new area of science was first mentioned in the 1960s as a subfield of statistics with a focus on "learning from data", also referred to as data analysis [41], and really took off around 2011.[3] Nowadays, topics such as artificial intelligence, machine learning, and data visualization, just to name a few, are or can be included under the umbrella term for computational data analysis that Data Science has become. The overall importance of Data Science from a Big Data perspective lies in i) its focus on learning from data in an evidence-based manner and ii) the efficient implementation of the algorithms[4] used.

One of the advantages of Data Science techniques is that they enable the automatic detection of structure (patterns) in large datasets. The previously mentioned study of website texts is an example of this. Here, computer algorithms are used to 'learn' (detect) the association between certain words and whether a business is innovative or not. In a similar way, associations can be extracted from other data sources, such as images. These approaches are often referred to as machine learning, i.e., the 'machine' (the algorithm) learns the association between certain features (words in this case) and an outcome. I will keep using the term 'machine learning' from here on, even though the more general term 'artificial intelligence' could also be applied. Data Scientists that apply machine learning techniques work in an inductive way. They search for patterns in data that can (hopefully) be generalized. This is the first step in developing a more general understanding of the phenomenon studied. How this can be done reliably with machine learning techniques in the context of official statistics is one of the topics we study. The ultimate goal of this work is to develop methodology that enables the reliable (trustworthy) use of machine learning techniques, e.g., machine learning methods, within official statistics.

So, what about the application of machine learning in a statistical context? Let me first say that machine learning is developing in an extraordinary fashion. Engineers are making giant leaps in this field, such as in the performance of the 'intelligent'

[3] See the Google Ngram Viewer [22] plot for the term 'Data Science'.

[4] In the context used here, an algorithm can refer to both an untrained and trained model.

ChatGPT chatbot I mentioned before. The application of machine learning has many links with statistics but tends to predominantly focus on the technique and on obtaining point estimates. There is less attention to bias and variance [42], topics that are very important for official statistics. Also, the transparency of findings based on machine learning algorithms is extremely important. Official statisticians want to understand how an algorithm comes to its outcome; we want to understand what is happening under the 'hood' of the algorithm. For example, a deep learning algorithm[5] can, when trained well, make decisions based on data with very high accuracy. But the essential question here is: do we understand how that algorithm comes to its (apparently good) decisions? If we do not understand it well enough, we have essentially developed a black box algorithm, which is not a good thing. I think that many of us are familiar with a recent example where an algorithm used by the Dutch tax office wrongly accused certain groups of people of social benefit fraud. This algorithm clearly had a bias and how it came to its 'recommendations' was certainly not transparent. Important topics that have been identified and are being studied in the context of Big Data are i) conceptualization, e.g., which variables (features) are included in the model and why, ii) the representativeness of the dataset used for training and testing, iii) the optimal sample size of the training set, and iv) the internal and external validation of the models developed.

I'll illustrate these topics with three examples. The first is an approach developed to identify online platform businesses based on the texts on their websites [26]. For this purpose, a model was developed based on a dataset containing known examples of both positive (platform) and negative (non-platform) cases. Because you need to start somewhere, experts from Statistics Netherlands were asked to provide a list of around 500 online platform businesses. The negative cases were obtained by taking an equal-sized random sample of all websites linked to businesses in the Business Register of Statistics Netherlands. The latter cases were thoroughly checked to ensure that no online platforms were accidentally included. Here, we assumed that the experts provided us with a representative sample of online platforms, but this does not have to be the case. One can imagine that these positive cases contain many more examples of businesses active in one (or more) particular branch(es). As a result, the model trained on those examples may miss specific groups of platform businesses active in other branches, which may have different features. Iteratively developing the model, extensive manual checking, and paying specific attention to the results obtained for various branches are ways to reduce this form of bias in the model. These and other potential approaches are being investigated [43].

[5] A class of machine learning algorithms that work based on the structure and function of the human brain, often described as multiple-layer artificial neural networks.

The second example is about internal and external validation. When one develops a model in a setting where there is a dataset with known outcomes (e.g., platform and non-platform cases), an 80% random sample is often drawn of such a dataset on which the model is subsequently trained. During training, the algorithm 'learns' the difference between the two cases (platform and non-platform) in the best possible way. The remaining 20% of the original dataset is used as an independent test set. This test set is used to independently determine how well the model is able to discern between the two cases as the test set contains examples that are entirely new to the model. We refer to this as the *internal* validation of the model's performance. However, for official statistics, we are predominantly interested in the performance of the model on the target population. In other words, for the online platform model, we want to know how well the model performs on totally new, unseen cases included in 'real-world' data. We refer to this as *external* validation. This requires data (if possible, with known outcomes) from a substantially larger dataset, ideally a representative part of the target population. A manual inspection of a sample of 'real-world' classified data by several experts is a way to determine the external validity of the model [44].

The third example has to do with the construction of a proper training and test set. In a supervised setting, a specific percentage of positive and negative examples are included in the dataset on which the model is trained. Quite often, 50% positive and 50% negative cases are used. However, these percentages may not have anything to do with the percentages to which these cases occur in 'real world' data. For instance, our best estimate of the percentage of online platforms in the Dutch Business Register suggests 0.25% positive cases [26, 43]. It is nearly impossible to train a useful model on a dataset with such a low number of positive cases simply because a model that always identifies a case as a non-platform will be correct in 99.75% of the cases. So, we need to use a different percentage of positive cases to obtain a model that is able to do that well, but what percentage is best? While looking at that, my colleague Marco Puts observed that a model trained on a particular percentage of positive items introduces a bias when applied to datasets with different (known) percentages of positive items [45]. His findings are shown in Figure 6.

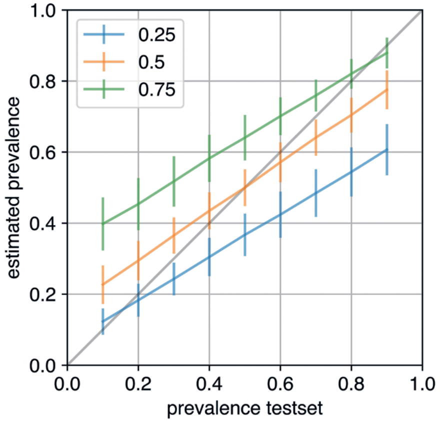


Figure 6. Effects of training a model on a certain ratio of positive items on the classification of datasets with different ratios of positive items [45].

Figure 6 reveals that models tend to be biased towards the outcome of the percentage of positives on which they are trained [45]. Along the x-axis, we see the *true* fraction of positive items, whereas the *estimated* fraction of positive items is shown along the y-axis. The gray line indicates the situation in which the true and estimated values are equal and the bias is thus zero. As one can see, the estimated value is only correct at one point: the fraction on which the algorithm was actually trained. Since online platforms are rare, there is a risk that a model developed on a dataset with increased prevalence will overestimate the number of online platforms when applied to 'real-world' data. This is actually what happened, but the effect was reduced by careful manual checking and validation of the outcome (of the model) by sending companies a questionnaire [26]. Currently, each step in this approach is being studied with the aim of improving it from the viewpoint of automating the selection process as much as possible. This work has resulted in the development of a new metric that can be used to improve the training of a model as it is less affected by high and low ratios of positive and negative examples [43]. It has also resulted in the development of a Bayesian adjustment method [46] to correct for the bias of a specific group of machine learning classifiers that produce (pseudo-)probabilities as their outcome.

An unexpected finding

We are now near the end of this lecture and it's time to illustrate the beauty of Big Data and data-driven findings as this provides the opportunity to discover unexpected associations. In 2017, me, my colleague Jan van de Brakel, and two others published a study on the correlation observed between the monthly Dutch Consumer Confidence Index and the average monthly sentiment of publicly available social media messages in the Netherlands [47]. The period studied was from June 2010 to March 2015. The Pearson correlation coefficient had an average value of 0.9 [48], which is very high, indicating a very similar development of both series. This was obviously not a spurious correlation as time series analysis studies confirmed that adding the social media sentiment series to the Consumer Confidence Index data improved the estimate; it reduced its variance [47]. This demonstrated that the sentiment series contained information on the concept measured by the Consumer Confidence survey. Looking upon this finding from a traditional statistical perspective, it is a very intriguing result. Let's apply the three important quality aspects mentioned at the beginning of this lecture concept, population, and stability over time to this example. Clearly, the concepts measured by both series, e.g., consumer confidence and the average sentiment of social media messages, are not identical, but there could be an overlap. Both could, for instance, be affected by an increase or decrease in the 'mood' of Dutch society [48]. The populations included in the data used, however, are clearly different. The survey is conducted on a representative sample of Dutch *households* [49] while the average sentiment is based on the sentiment of all public *messages* sent by Dutch people active on social media during each month [47]. So, we are comparing households and messages here as the base units! We also need to realize that the people active on social media differ considerably from the average Dutch person [49, 50] and that the number of messages sent may differ considerably over time [48]. Hence, it is remarkable that a correlation was found and that it is apparently clearly associated with consumer confidence. In that respect, the third quality criterion, stability over time, is very positive in this case. A recent study, performed at the end of 2022, revealed that the association is still there [51]. I hope that you all agree with me that Big Data is an intriguing topic indeed.

Concluding remarks

I'm almost at the end of my lecture. What I have been trying to clarify for you is that the use of Big Data in official statistics is a broad topic for which many things still need to be figured out. This requires the need for research in the areas of i) the quality of Big Data, with special attention to the (internal and external) validation of concepts, producing accurate Big Data population-based estimates, and dealing with concept drift, ii) combining Big Data with other sources, iii) enabling the reliable use of Data Science (machine learning) methods within an official statistical context, and iv) studying various potential applications in a data-driven way. All of this work is within the scope of my chair, with the ultimate goal to develop a generalizable Big Data methodology. That is quite a challenge. Luckily, I collaborate with very good statisticians and data scientists at Statistics Netherlands, have a very good PhD involved, am regularly supported by good students (from various universities, incl. TU/e) during their internship, and am in contact with world-renowned experts at Eindhoven University of Technology and other institutes worldwide. This in addition to living in the smartest region of the world! is a privilege indeed.

Oh, for the record, the term 'Big Data' has been used 91 (or 92?) times in this document so far.

Acknowledgments

We have come to the end of my speech. I would first of all like to thank my partner in life, Marijke, and our three kids, Quinty, Thijmen, and Sterre. Marijke, thanks for your support over the years during my academic, non-academic, and - again - academic periods. It has been quite a journey. It's great having you in my life, not only as a partner but also as somebody to laugh with and to ensure that I keep my feet on the ground. I can always count on you, even when I have to travel abroad and forget to inform you on time. Quinty, Thijmen, and Sterre, I'm blessed to have you in my life. I'm very proud that you all study or have studied at university (not in Eindhoven unfortunately) and I hope you continue to do well and enjoy life as much as you can. I'm grateful to my mother and father, who raised me well (in a toy shop) and gave me the time to find out what I should be doing in life. It took me a while to figure out that I was actually a scientist.

Next, I thank Statistics Netherlands and Eindhoven University of Technology for providing me with this challenging scientific opportunity. I'm especially grateful to the Rector Magnificus and to Dean and Professor Edwin van den Heuvel of Eindhoven University of Technology for enabling this. From the Statistics Netherlands side, I specifically thank Jacobiene van der Hoeven, Sofie de Broe, and former Director General Tjark Tjin-a-Tsoi for their contributions and support. Without all of them, my chair would not have been created. I also thank Vera Toepoel and Joost Huurman for their ever so subtle encouragement to give this lecture (post-COVID, of course).

A number of people have been, and still are, extremely important and dear to me. First and foremost, I would like to thank Marco Puts. Without you, the Big Data work presented wouldn't be half as good, would be less innovative, and certainly less fun. I really enjoyed our brainstorming and hacking sessions in the early days of Big Data (when the size of the Big Data group was two), our highly frequent joint presentations for visitors from abroad (the legendary presentation on “finding a needle in a haystack” comes to mind), and our lunches at Subway in Heerlen. That's actually where the whole 'Big Data methodology' journey started! Thanks, Marco, for your support, creativity and unfiltered comments, and for being a friend. Next, I want to thank 'my' PhD colleague Yvonne Gootzen. I'm grateful to have you on board and really enjoy working with you. If the speed at which your first paper

was accepted is an indication of the future, your scientific career will skyrocket. I would also like to thank all data scientists/statisticians active in the Big Data, Data Mining, and AI research program of Statistics Netherlands, viz. Martijn, Marc, Chris, Joep, Jonas, and Tim, and all statisticians of the Statistics group of TU/e. I also thank all of my NS travel colleagues on the Eindhoven-Heerlen trajectory for their enjoyable company (especially during train failures), my methodological colleagues in Heerlen and the Hague, anyone working at Statistics Netherlands with Data Science and/or Big Data at heart, and all my international Big Data colleagues.

Last but certainly not least, I thank you, the audience, for being here, including all of the people connected online, for listening to this lecture. I especially like to thank all of 'my' students (starting with Beike Hendriks in Nijmegen in 1990 up to Sanne Peereboom from Leiden this year) for their contributions. I learned a lot from you and, hopefully, you also learned a lot from me. Again, thank you for your attention and that's it.

Dixi.

References

1. United Nations (2014). *Fundamental Principles of Official Statistics*. United Nations Statistics Division. https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx
2. Bethlehem (2009). *Applied Survey Methods: A Statistical Perspective*. Wiley.
3. Wallgren and Wallgren (2014). *Register-based statistics: Statistical methods for Administrative Data*, Wiley.
4. Bethlehem (2009). The rise of survey sampling. Discussion paper 09015, Statistics Netherlands. https://www.cbs.nl/-/media/imported/documents/2009/07/2009-15-x10-pub.pdf
5. Schulte Nordholt (2018). "The usability of administrative data for register-based censuses". *Stat. J. IAOS* 34(4), 487-498.
6. Mayer-Schönberger and Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. John Murray.
7. Pentland (2015). *Social Physics: How social networks can make us smarter*. Penguin Books, page x.
8. Roos, Daas, and Puts (2009). Innovative data collection: New sources and opportunities. Discussion paper 09027, Statistics Netherlands. https://www.cbs.nl/nl-nl/achtergrond/2009/20/waarnemingsinnovatie-nieuwe-bronnen-en-mogelijkheden (in Dutch)
9. Knuth (1962). *The structure of scientific revolutions*. Univ. Chicago Press.
10. De Broe et al. (2020). "Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics". *Stat. J. IAOS* 37(1), 343-360.
11. Statistics Netherlands (2023). The statistical process. https://www.cbs.nl/en-gb/about-us/organisation/the-statistical-process
12. Kitchin (2015). "The opportunities, challenges and risk of big data for official statistics". *Stat. J. IAOS* 31(3), 471-481.
13. Struijs, Braaksma, and Daas (2014). "Official Statistics and Big Data". Big Data Soc. 1(1), 1-6.
14. Leskovec, Rajaraman, and Ullman (2020) *Mining of Massive Datasets*. 3rd edition. Cambridge Univ. Press.
15. Burnaev et al. (2021). *Analysis of Images, Social Networks and Texts*. Revised selected papers of the 10th International Conference, AIST 2021. Springer.

16. Di Bucchianico et al. (2019). Mathematics for Big Data. In Pitici (Ed.), *The Best Writing on Mathematics 2019* (pp. 120-131), Princeton Univ. Press.
17. Adler and Rips (2008). *Reasoning: studies of human inference and its foundations*. Cambridge Univ. Press, Part II: Modes of reasoning.
18. Coveney, Dougherty, and Highfield. (2016). "Big data need big theory too". *Phil. Trans. R. Soc. A.* 374, 20160153.
19. Diebold (2012) "On the Origin(s) and Development of the Term 'Big Data'". PIER Working Paper No. 12-037, http://dx.doi.org/10.2139/ssrn.2152421
20. Google (2023). Google search website. https://www.google.com/
21. ChatGPT (2023). ChatGPT website. https://chat.openai.com/
22. Google (2023) Books Ngram Viewer website. https://books.google.com/ngrams/
23. Puts, Daas, and de Waal (2017). Finding Errors in Big Data. In Pitici (Ed.), *The Best Writing on Mathematics 2016* (pp. 291-299), Princeton Univ. Press.
24. Gootzen, Daas, and van Delden (2023). "Quality Framework for combining survey, administrative and big data for official statistics". *Stat. J. IAOS*. https://doi.org/10.3233/SJI-220110
25. Fretwell and Tratan (2021). "Discovery of new colonies by Sentinel2 reveals good and bad news for emperor penguins". *Remote. Sens. Ecol. Conserv.* 7(2), 139-153.
26. Daas, Hassink, and Klijs (2023). "On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms". *Submitted for publication*.
27. Gebru et al. (2017). "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States". *PNAS* 114(50), 13108-13113.
28. Daas and van der Doef (2020). "Detecting Innovative Companies via their Website". *Stat. J. IAOS* 36(4), 1239-1251.
29. Soldaat et al. (2017). "A Monte Carlo method to account for sampling error in multi-species Indicators". *Ecol. Indic.* 81, 340-347.
30. Schiavoni et al. (2021). "A dynamic factor model approach to incorporate Big Data in state space models for official statistics". *J. R. Stat. Soc. Ser. A.* 184(1), 324-353.
31. Kennedy-Martin et al. (2015). "A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results". *Trials* 16, 495.
32. Callaci et al. (2019). The tradeoff between the utility and risk of location data and implications for public good. Connected Life 2019 conference, Oxford. https://arxiv.org/abs/1905.09350

33. Meng (2018). "Statistical Paradises and Paradoxes in Big Data (I): Law of Large populations, Big Data Paradox, and the 2016 US Presidential Elections". *Ann. Appl. Stat.* 12(2), 685-726.
34. Wu (2022). "Statistical inference with non-probability survey samples". *Surv. Methodol.* 48(2), 283-311.
35. Beręsewicz et al. (2018). An overview of methods for treating selectivity in big data sources. Statistical working papers, Eurostat publications. https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf
36. Axenbeckl and Breithaupt (2021). "Innovation indicators based on firm websites – Which website characteristics predict firm-level innovation activity?" *PLoS One*, 0249583.
37. Daas and Jansen (2020). "Model degradation in web derived text-based models". Paper for the 3rd International Conference on Advanced Research Methods and Analytics (CARMA), Valencia/online. http://dx.doi.org/10.4995/CARMA2020.2020.11560
38. Ginsberg et al. (2009). "Detecting influenza epidemics using search engine query data". *Nature* 457(7232), 1012–1014.
39. Lazer et al. (2014). "The Parable of Google Flu: Traps in Big Data Analysis". *Science* 343 (6176), 1203–1205.
40. Bostanci, Gootzen, and Lugtig (2023). Data Linkage to Validate and Calibrate Traffic Estimations on a Nationwide Scale: A Framework for Official Statistics. Discussion paper, Statistics Netherlands. https://www.cbs.nl/-/media/_pdf/2023/11/discussion-paper-bostanci-2023.pdf
41. Donoho (2017). "50 Years of Data Science". *J. Comp. Graph. Stat.* 26(4), 745-766.
42. Puts and Daas (2021). "Machine Learning from the perspective of Official Statistics". *The Survey Statistician* 84, 12-17.
43. Gubbels (2023). The sample Pearson Correlation Coefficient for classification models and identifying platform economy businesses from web-scraped data. Master's thesis Applied and Industrial Mathematics, Eindhoven University of Technology, the Netherlands.
44. Daas, De Miguel, and De Miguel (2023). "Identifying Drone Web Sites in Multiple Countries and Languages with a Single Model". *J. Data Sci.* 23-JDS1087.
45. Puts and Daas (2021). "Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach". Symposium on Data Science and Statistics (SDSS) 2021, online. https://arxiv.org/abs/2102.08659
46. Puts (2023). BayesCCal: Bayesian Calibration of classifiers, Code on Github. https://github.com/mputs/BayesCCal

47. Van den Brakel et al. (2017). "Social media as a data source for official statistics; the Dutch Consumer Confidence Index". *Surv. Methodol.* 43(2), 183-210.
48. Daas and Puts (2014) Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series no. 5, Frankfurt.
49. Constantinides et al. (2010). Profiles of Social Networking Sites Users in the Netherlands. In: The 18th Annual High Technology Small Firms Conference, University of Twente, the Netherlands.
50. Statistics Netherlands (2020). Who uses social media the most? https://longreads.cbs.nl/nederland-in-cijfers-2020/wie-gebruikt-het-vaakst-sociale-media/ (in Dutch).
51. Daas et al. (2022) Validation of the Social Unrest Indicator. Statistics Netherlands report. https://www.cbs.nl/-/media/innovatie/validatie_soi.pdf (in Dutch).

Appendices

APPENDIX A: PRIVACY CONSIDERATIONS AND BIG DATA

Sometimes, Big Data sources contain personal data. The General Data Protection Regulation (GDPR)[6], a European regulation that helps to protect the privacy of citizens, defines 'personal data' as any information relating to an identified or identifiable natural person ('data subject'). Personal data also include data that indirectly reveal something about a natural person. Examples of personal data are a name, a home address, an email address, or someone's location.[7] When studying data sources containing personal data, it is essential to comply with the GDPR regulation. Both Eindhoven University of Technology (TU/e)[8] and Statistics Netherlands (CBS)[9] adhere to the GDPR and its Dutch Implementing Act (AVG)[10]. Apart from the GDPR, Statistics Netherlands additionally adheres to the privacy stipulations in the Statistics Netherlands Act[11], the European Statistics Code of Practice[12], and its own Code of Conduct.[13] As such, Statistics Netherlands meets the most stringent requirements regarding data protection. In addition, an accredited external organization performs a yearly privacy audit. Article 44 of the Dutch GDPR Implementing Act[x] includes exceptions concerning scientific and statistical research.[14] This article enables employees of institutions that perform such research, such as employees of TU/e and Statistics Netherlands, to carry out research and/or produce statistics from data including personal data *provided that necessary (privacy provision) measures have been taken*.

[6] GDPR (2023). General Data Protection Regulation. https://gdpr-info.eu/

[7] GDPR (2023). What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr/

[8] TU/e (2023). Privacy and ethics. https://www.tue.nl/universiteit/library/library-for-researchers-and-phds/research-data-management/rdm-themes/privacy-and-ethics/

[9] Statistics Netherlands (2023). Privacy, CBS. https://www.cbs.nl/en-gb/about-us/organisation/privacy

[10] AVG (2023). "De Algemene Verordening Gegevensbescherming". https://wetten.overheid.nl/BWBR0040940/2021-07-01 (in Dutch).

[11] Statistics Netherlands (2023). Statistics Netherlands Act. https://www.cbs.nl/-/media/_pdf/2017/28/statistics-netherlands-act-2019.pdf

[12] Eurostat (2023). European Statistics Code of Practice. https://ec.europa.eu/eurostat/en/web/products-catalogues/-/KS-02-18-142

[13] Statistics Netherlands (2023). "Gedragscode". https://www.cbs.nl/-/media/cbs/over-ons/organisatie/gedragscode.pdf (in Dutch).

[14] AVG (2023). "Uitzonderingen inzake wetenschappelijk onderzoek en statistiek", Artikel 44 van de AVG. https://wetten.overheid.nl/jci1.3:c:BWBR0040940&hoofdstuk=4&artikel=44 (in Dutch).

The following measures are generally recommended when working with data sources that contain personal data:

Work safely: Do not leave printouts on the printer or desk, do not use public Wi-Fi, do not work where others can easily watch your screen or hear you talk, do not leave your laptop logged in when you are away, etc.

Purpose limitation: Only use personal data for specified, explicit and legitimate purposes.

Informed consent: Make sure that your data subjects are well informed about the purpose of the research, the risks, and the data processed. Ask them to sign an informed consent form.[15,16]

Data minimization: Only process personal data that are relevant and necessary for the purpose of your research. Thus, include as little personal information as possible.

Storage (limitation): Store personal data in a form that permits identification of the data subject for no longer than is necessary for the purposes of your research. Store identifiable information apart from other information. De-identified data should be kept for a period appropriate to the respective discipline and methodology.

Work with de-identified data: The best way to protect your data subject's privacy is to not collect certain identifiable information at all. Otherwise, data that can identify a person with little to no effort needs to be de-identified. Personal data can either be anonymized (data cannot be traced back to an individual person) or pseudonymized (no immediate identification but it remains possible to identify a person from the data with the use of additional information).

Anonymization: Remove all direct identifiers (name, address, telephone number, etc.) but also indirect identifiers (age, place of birth, occupation, salary, etc.) that, linked with other information, can lead to a person's identification. Be aware that a combination of indirect identifiers might enable one to identify a data subject. Anonymization to the point that a data subject is no longer identifiable means that the anonymized data is not considered to be personal data anymore.

Pseudonymization: Replacing the unique identifier of a data subject with an artificial pseudonym. This means that identification is still possible with the identification key. The identification key needs to be stored securely and separately from the pseudonymized data. If the data subject can be identified by combining data with additional information, the data is also called pseudonymous.

Encryption: If it is not feasible to de-identify the data, encrypting data is an option. Encryption makes data unreadable or inaccessible to those without a password. The data file itself can be encrypted or the hard drive, where the data file is located, can be encrypted.

Restricted access: Define who has access to your data, how you will restrict this, how you will enable access to those authorized, and where you will describe who gets access to the data.

Processing agreement: This is a contract in which you lay down the responsibilities both parties have concerning the processing of personal data. Topics that are covered in a processing agreement are a general description of the project, confidentiality, security, compliance with the GDPR, privacy rights, etc.

Article 5(1b) and Article 89(1) of the GDPR and Article 44 of its Dutch Implementing Act enable research and statistics production from data including personal data. Article 9 of the GDPR and Article 24 of its Dutch Implementing Act describe regulations regarding the processing of special categories of personal data for these purposes. Without those articles, research and statistics production would *not* be possible. Let it be clear that all of the research described in this document has been conducted in accordance with the GDPR regulations.

[15] For Big Data this is usually not possible. Here, anonymization is the recommended option.

[16] Viganò et al. (2022). Ethical, Legal and Social Issues of Big Data - A Comprehensive Overview. White Paper of the ELSI Task Force for the National Research Programme "Big Data". https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4081192.

APPENDIX B: HARDWARE, SOFTWARE AND BIG DATA

Processing large amounts of data greatly benefits from an efficient infrastructure (hardware) and effectively implemented programming languages (software). Both are essential when studying Big Data as they will ensure that not too much time is wasted while waiting for the results.

Hardware

Regarding the hardware, processing large amounts of data in as short time intervals as possible is the goal. Hence, the need arises for the availability of more computational power. In principle, two approaches can be applied to enable this. The first is known as vertical scaling ('scaling up') and the second is known as horizontal scaling ('scaling out').

In the case of vertical scaling, the increasing demand for computing power is dealt with by adding additional resources to an existing system or by buying a completely new 'bulkier' system. Here, existing hardware is replaced by faster alternatives, i.e., the computer is scaled up. Vertical scaling works best when one studies topics that can't be easily split into subtasks that can be processed in a parallel way. For instance, when there are interdependencies (requiring the exchange of intermediate results) between parts of a distributed dataset. Also, many iterative methods are inherently serial.

In the case of horizontal scaling, increasing computational demands are met by adding additional computers to an existing infrastructure. Here, the system is scaled out. Because more computers are available, the data can be distributed over more units, which (usually) speeds up the analysis. Be aware that the results need to be combined at the end of the analysis. Horizontal scaling works best when studying problems that can be easily parallelized, such as counting words in individual sections of documents. Such tasks are commonly referred to as embarrassingly parallel.

Although not written very recently, the 2015 paper of Singh and Reddy[17] provides an excellent overview of the various vertical and horizontal scaling platforms one can use. For horizontal scaling, Spark is (at the time of writing) generally considered the best framework. For vertical scaling, there is no single best solution. Depending on the problem studied, High Performance Computing clusters, multicore processors, Graphics Processing Units (GPU), and Field Programmable Gate Arrays (FPGA) may be valid options.

Software: programming languages

A number of computer programming languages can be used for the analysis of Big Data. The most often mentioned open sources languages are R, Python, Julia, and Scala. Not co-incidentally, the same programming languages are also preferred by many data scientists.

R is a programming language created by (bio)statisticians.[18] It is often used for exploratory data analysis. R has a very active community that has created an extensive set of R-packages[19] that enable it to integrate with many Big Data analytics environments. It also has great visualization capabilities. It is not the fastest language around but, by making extensive use of libraries implemented in C or C++, this downside can be considerably reduced.

Python is a general purpose programming language.[20] The conciseness of its syntax has greatly stimulated its success. In the context of Big Data, it is often the language of choice for machine- and deep learning and for natural language processing. Also, many additional libraries are available for Python, which can be easily added.[21] Python is not the fastest language around but is a very good and generally applicable choice.

Julia is a programming language designed from the beginning for performance.[22] This makes it a very attractive language for Big Data analysis as it is usually much faster than the alternative options. Compared to R and Python, not as many libraries are available in the Julia package archive[23], but this number is steadily growing. If this growth continues, it could make it the Big Data analytics language of the future.

[17] Singh and Reddy (2015). "A survey on platforms for big data analytics". *J. Big Data* 2(8), s40537-014-0008-6.

[18] R (2023). The R project for Statistical Computing, R. https://www.r-project.org/

[19] CRAN (2023). Comprehensive R Archive Network https://cran.r-project.org/

[20] Python (2023). Python programming language. https://www.python.org/

[21] Python libraries (2023). The Python Package Index. https://pypi.org/ & CONDA website, https://docs.conda.io/en/latest/

[22] Julia (2023). Julia programming language. https://julialang.org/

[23] Julia archive (2023). Julia package archive. https://juliapackages.com/

Scala stands for 'scalable language' and is a programming language specific for the analysis of Big Data on distributed systems.[24] The language enables one to define a program that can be run in serial, in parallel across all available cores on a single machine, or in parallel across a cluster of machines without changing the code. It is therefore considered by many the language of choice for the analysis of huge amounts of data on distributed systems. There is an 'awesome' website with an overview of Scala libraries.[25]

Using other languages for Big Data analysis, such as SAS, Java, and C/C++, is not excluded, of course.

[24] Scala (2023). Scala programming language. https://www.scala-lang.org/

[25] Awesome Scala (2023). Awesome Scala website. https://index.scala-lang.org/awesome

Curriculum Vitae

Prof.dr. Piet J.H. Daas was appointed part-time professor of Big Data in Official Statistics at the Department of Mathematics and Computer Science at Eindhoven University of Technology (TU/e) on January 1, 2019.

Piet Daas (1963) obtained his MSc in Biology (1990) and PhD in the Natural Sciences (1996), both cum laude, at Radboud University in Nijmegen. After a postdoc period at the Wageningen University Laboratory of Food Chemistry (1996-2000), he started working at Statistics Netherlands ('Centraal Bureau voor de Statistiek'). Here, he became a senior methodologist at the Department of Data Services, Research and Innovation and an expert in the (re-)use of existing data for statistical purposes. He has been performing pioneering studies on the use of Big Data for official statistics since 2007 and has been the project leader of the corresponding research theme since 2011. He is a renowned Big Data expert and has been involved in various European, UN/UNECE, and US National Academies of Sciences initiatives. He is an instructor for a number of Big Data courses at the Statistics Netherlands Academy, the Eurostat European Statistical Training Program, the European Masters of Official Statistics of Utrecht University, and the Advanced Topics in Official Statistics course of the universities of Mannheim and Munich.

Colophon

Production
Communication Expertise
Center

Cover photography
Bart van Overbeeke
Photography, Eindhoven

Design
Grefo Prepress,
Eindhoven

Digital version:
www.tue.nl/lectures/

Visiting address
Building 1, Auditorium
Groene Loper, Eindhoven
The Netherlands

Navigation
De Zaale, Eindhoven

Postal address
PO Box 513
5600 MB Eindhoven
The Netherlands
Tel. +31 (0)40 247 9111
www.tue.nl/map

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY