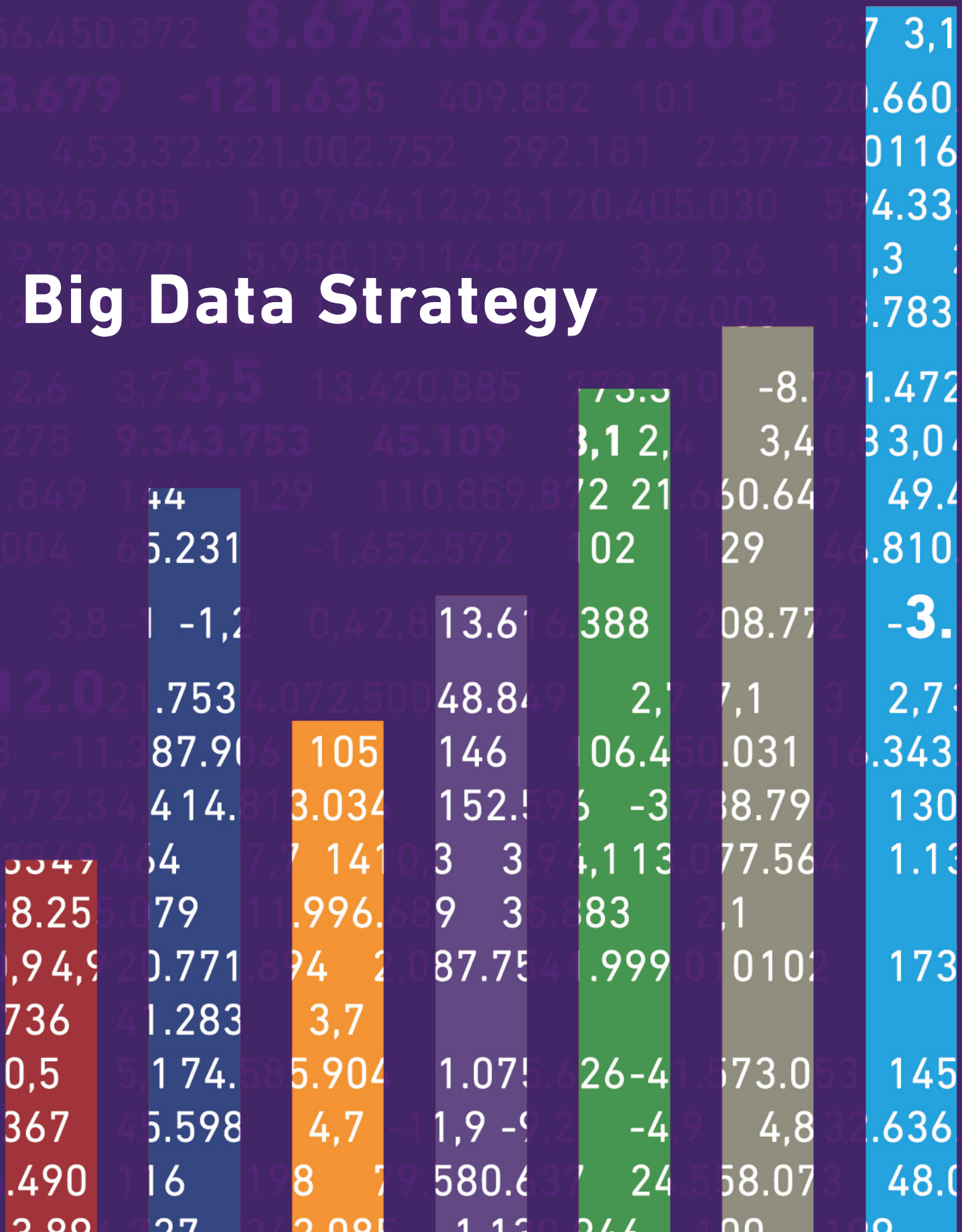
STATISTICS DENMARK

Big Data Strategy



Big Data Strategy 2018-2020

STATISTICS DENMARK

Big Data Strategy 2018-2020

Published by Danmarks Statistik
October 2018
Photo: Graphics by Statistic Denmark

Pdf-version:
Downloadable at www.dst.dk/publ/BigDataStratUK
ISBN 978-87-501-2306-4

Adresses:
Statistics Denmark
Sejrøgade 11
2100 Copenhagen Ø, Denmark

Tel. +45 39 17 39 17

e-mail: dst@dst.dk
www.dst.dk

Introduction

Big Data is the effect of the accelerating digitisation, which means that individuals as well as enterprises leave behind electronic footprints while performing a vast number of more or less standard and everyday actions. This happens via online activities (shopping in a very wide sense, transport, social media, and media consumption in general and registration of personal activities, e.g. health and sporting activities) and business transactions (purchase, sale of goods and services, and goods transport). The use of digital sensors in meters (e.g. electricity meters) and machines in the broad sense (e.g. means of transportation and agricultural machinery) (the so-called internet of things) is another wave in the creation of Big Data.

Big Data sources distinguish themselves from other known data sources (administrative data and survey data) in a number of ways, in terms of quantity (large), sources (multiple), speed (high — data generation is continuous), variation (in sources and their structure), and reliability (of data sources in relation to a given purpose).

Big Data may be (part of) the answer to a number of challenges faced by official statistics, such as declining response rates for traditional surveys, reduced financial resources and the request for timelier estimates.

Statistics Denmark's Strategy 2022 stipulates that an action plan for utilisation of Big Data must be prepared in the strategy period, and that partnerships must be set up with producers of Big Data on the application of such data in the statistical production (Statistics Denmark, Strategy 2022, p. 14).

Over the 2018-20 period, Statistics Denmark's Big Data Strategy will focus mainly on the application of data in relation to existing statistics and on forming data partnerships with others for the purpose of improving existing sources for the official statistics.

This strategy contributes towards this and describes the strategic commitment to Big Data in the following ways:

- Experience with Big Data at Statistics Denmark.
- Competences and competence development
- The legal aspects
- Partnerships
- International cooperation
- And how these efforts should be organised

Experience with Big Data at Statistics Denmark

Statistics Denmark already has some experience in the use of Big Data.

The use of *bar code data* as input for the consumer price index is the only example of the use of Big Data in the production of statistics. An example that demonstrates the need for thorough investigations of the challenges associated with the use of Big Data in the production of statistics. The work was set in motion in 2010 and, as a result, we incorporated bar code data in the production of the consumer price index from 1 January 2016. In this period, we conducted thorough investigations and drew on the experience of other countries before the model for application of the new data was in place.

AIS data (Automatic Identification System) are digital notifications about the positions of all ships in Danish waters[1]. Statistics Denmark has participated in a working group under ESSnet Big Data, where we studied the application of AIS data. AIS data could be e.g. a supplementary data source for Passenger and ferry traffic, and be input for the Green National Accounts (emissions of CO2 and NOx) and Tourism statistics.

Electrical meter data from energinet.dk. In 2020, all electrical meters in Denmark (and the rest of the EU) must be smart meters that transmit the consumption to the supplier every 15 minutes. This allows us to monitor the consumption of electricity very closely. Statistics Denmark participates in a working group under ESSnet Big Data, where we study the application of SmartMeter data. Electrical meter data can be e.g. the data source for a more detailed set of energy statistics as well as housing statistics.

Web scraping – collection of data directly from the internet. In 2016, Statistics Denmark examined the possibility of integrating web scraping in the data collection for the statistics for job vacancies. Back then, a number of problems came up which may since have been solved. As an example, Eurostat's ESSnet Big Data has focused their efforts in recent years on collecting data from the internet and tried to solve the problems they have encountered in the process. Web scraping is used mainly as a source of quality assurance of data when the primary data comes from another source.

Payment card data – contains payment transactions carried out either in physical ATMs or online. An important purpose of accessing payment card data for Statistics Denmark is to improve the debit side in the balance of payments' travel item, since this is a notoriously difficult item to calculate. However, there is a broad range of potential applications of payment card data. So far, we have managed to gain access to a test data set that covers a period of 18 months.

Moreover, a number of initiatives exist in relation to reporting business statistics, in particular about the use of *Digital accountant* for the accounts statistics, *Sensor data* for transport and agricultural statistics, *data from the platform economy* as further automation of the enterprises' reporting of data to Statistics Denmark. In addition, there are further Big Data sources that we can explore in more detail.

[1] All ships above a certain size (>300 gross tonnes, all passenger ships and all fishing boats of a length over 15 meters) must be equipped with an AIS transponder.

Activities

- Make an analysis of current and future use of Big Data sources for existing and new official statistics, including a conceptualisation of various forms of application of Big Data sources – e.g. in replacement of existing data sources, as a supplement to existing data sources, as a basis for early estimates or new statistics

- Expand the application of *bar code data* for full coverage of supermarket chains and to cover other areas, e.g. filling stations, and to perform an analysis of other potential applications

- Testing the application of *AIS-data* in one or more specific areas of statistics, e.g. green national accounts and statistics of harbours

- Enhance the experience with the use of *electrical meter data* in preparation for application in a concrete production of statistics with data from 2020 upwards, e.g. the housing statistics

- Revisit the potential for using *web scraping* for quality assurance of information about job vacancies from the public sector bringing in the experience of other statistical institutions, e.g. Statistics Netherlands and Republic of Slovenia Statistical Office.

- Uncover the potential for using *payment card data* in a number of existing and new fields of statistics, such as the balance of payments' travel item and e-commerce, by means of the existing test data

- Uncover the potential for increased automation of the reporting by private enterprises, including the reports submitted by farms to Statistics Denmark through voluntary data partnerships with the suppliers of the system solutions for the private sector

- Examine the potential for use of further data sources, e.g. mobile phone data and data (e.g. travel card) on public transport for "congestion statistics", data from social media for statistics on "public sentiment" and the field database for agricultural statistics.

Competences and competence development

Statistics Denmark's process model[2] applies most directly to traditional survey-based statistics, but can also be applied to describe statistics based on Big Data.

The competence requirements in relation to Big Data are about how to get from raw data in often unknown or laborious formats to useful facts and knowledge that can be extracted and disseminated. In addition to the strictly technical challenges, extensive knowledge is also required about the applicability of data, since the data generating process is rarely well described. E.g., it is not trivial to describe a relevant sample.

As for the practical competence development requirement in Statistics Denmark, we are working with roles and competence levels in order to address various types of employees.

There are two main roles: Heads of statistics/statistical staff as well as IT staff and employees of the IT department. A secondary role is the management, who must have a basic level of appreciation to be able to approach the Big Data potential at the operational level.

We mainly handle the competence development in relation to Big Data internally. Competent employees in IT as well as in Methodology and Analysis will provide relevant education and training supplemented with external resources for a few select areas (such as machine learning).

Cross-functional collaboration is a necessary skill and discipline for a Big Data project to succeed. In particular, before a specific data area based on Big Data is operationally mature and ready for publishing or dissemination, it will require collaboration across the statistical sections, Methodology and IT, focusing on IT competences as well as competences in the statistical sections for interpretation of data from Big Data sources.

Moreover, internal as well as external networking is relevant and valuable. It is highly rewarding to keep up with the progress of other statistical institutions in the field of Big Data and the dialogue with Nordic and international colleagues provide a basis for valuable sparring and knowledge.

We estimate that the R environment together with the existing Oracle and SAS platforms cover the need for tools and technological capacity for the moment.
New tools and capacity must primarily be driven by concrete needs together with the general considerations that exist in Methodology and IT about the tools portfolio of the future.

[2] An adaptation of the generic GSBPM – Generic Statistical Business Process Model

Activities

- Create an overview of the specific tools and the development in the use of these tools applied by the statistical institutions that are most advanced in using Big Data for official statistics

- Provide an overview of competences in the statistical sections and IT department and make a plan for competence development in the use of Big Data for official statistics

- Elaborate on the competence development plans in Statistics Denmark's revised IT strategy focusing on competence requirements to deliver on the challenge of using Big Data for the production of official statistics

- Uncover the competence requirements in relation to the development and use of new ways of storing very large volumes of data

- Organise competence partnerships across the organisation and externally (digital task force) to quickly (sprint) develop and test ideas in strategically selected areas of action

The legal aspects

This section deals with the possibilities of introducing legislation that makes it mandatory for private enterprises (data owners) to submit Big Data to Statistics Denmark for statistical purposes.

In this context, there are two tracks. We can make efforts to obtain national legislation securing this at the national level or to obtain an EU regulation applying to all member states – a combination of these will also be an option.

If we make efforts to obtain a solution via national legislation, we can carry on from the basis of the preparatory work made in connection with the amendment of the Act on Statistics Denmark.

However, the possibilities via national legislation have been put on hold after the act was adopted without any provision of authority to impose requirements in connection with the collection of Big Data.

For this reason, the logical track to pursue is via an EU legislative act, a regulation (which has immediate effect in all member states), a directive (which must be implemented in the member states' own legislation) or via EU soft law, which is not immediately legally binding (e.g. various forms of agreements).

In terms of using the statutory instrument to make reporting of Big Data for statistical purposes mandatory, we must take into account that since 2017, there has been a strong change in the general public concern for the security of the citizens' digital data – with public as well as private data owners. This is a result of e.g. the discussions in relation to the implementation of the EU General Data Protection Regulation and examples of unsafe handling of the citizens' data by public authorities.

In this way, the recommendation in connection with any legislative initiative could be to provide a guarantee that the Big Data collected by the national statistical institutions for statistical purposes is of a nature that does not allow it to be used for any other purpose – not even if it is disclosed illegally, and that the research application is limited to prevent any situations of abuse. This means that the rather broad phrasing in section 10 (2) of the Danish Data Protection Act, where it says that data processed for the sole purpose of carrying out statistical or scientific studies may not subsequently be processed for any other purposes. This means that it must appear from the legislation under consideration how to secure this narrow use.

Activities

- Prepare a note based on international experience about conditions involving the legal aspects of the access by official statistics to use Big Data for the production of official statistics

- Participate in the development concerning access to Big Data for statistical purposes in the international statistical system (Eurostat and the UN)

- Participate actively in the debate with discussions and events showing specific examples of the socio-economic potential of securing statistical and research institutions' access to Big Data in ways that keep the data of citizens and enterprises secure.

Partnerships

As specified, Statistics Denmark has entered into a partnership with a number of supermarket chains on reception and use of scanner data as input for the consumer price index. In addition to this, there have been sporadic contacts with other potential suppliers of Big Data, e.g. in the field of mobile telephony.

We have also been in contact with trade organisations (the Confederation of Danish Industry and the Danish Chamber of Commerce) about Big Data and with the academic world (University of Copenhagen, the Technical University of Denmark and the IT University of Copenhagen), e.g. about the creation of a master programme in data science as a graduate programme for the Bachelor of Science programmes at the University of Copenhagen.

Finally, we have been in contact with Microsoft Denmark about the use of Big Data, just as we have been in contact with Microsoft Seattle who have also paid us a visit, and we have discussed potential collaborative projects.

As part of its strategy, the UN Global Working Group on Big Data for official statistics has developed good relations with big global tech corporations (Microsoft, Google, Amazon and Nielsen) in preparation for joining forces on Big Data projects that are useful and relevant for the compilation of official statistics by using Big Data.

There is hardly any doubt that regardless whether we obtain access to use Big Data for the production of official statistics via legislation or not, the development in this field must happen in the form of partnerships.

Partnerships may have multiple purposes in addition to delivery of Big Data for official statistics. It could be partnerships involving mutual exchange of data, where data from the private data supplier is included in Statistics Denmark's production of official statistics, and where data from Statistics Denmark is used to improve the data sources of a private data supplier. It could also be partnerships involving competence development and data storage, and partnerships that allow Statistics Denmark to gain an insight into and become a partner in the development of new ways of using data with the private data supplier; and partnerships where Statistics Denmark makes data available to researchers and analysts, as is the case with data from Energinet.

Forming partnerships concerning data sources as well as technology and competence development is a central part of the strategy for this area. This approach is supported by the endeavours of the UN Global Working Group to form global partnerships in support of the efforts made by the individual statistical institutions in this area.

In Denmark, it also possible to try to form data partnerships with suppliers of "data reporting systems" for the private sector, and with trade organisations on using new ways to report data.

Activities

- We will prepare a presentation which shows potential partners and the potential of Statistics Denmark as a Big Data partner, focusing on our special strong points in relation to data application, data documentation and data sharing

- We will establish a system for the contact with potential Big Data suppliers (including the Danish trade organisations) in preparation for a dialogue on the potential of using Big Data for official statistics

- We will establish a system for the contact with academia and the tech corporations with a survey of the potential in setting up an advisory board for the use of Big Data for official statistics with the participation of e.g. the University of Copenhagen, the Technical University of Denmark, the IT University of Copenhagen and Danish tech corporations

- We will carry out a session of interviews (tech lunches) with selected Big Data suppliers, academia and tech corporations as part of the planning of Statistics Denmark's work programme for 2019

- We will examine and be specific about the possibility of forming voluntary partnerships with system suppliers in selected areas

- We will examine the possibility of carrying through projects with Microsoft in relation to Statistics Denmark's work programme for 2019.

International cooperation

At its 45th session in 2014, the United Nations Statistical Commission set up a working group to promote the use of Big Data for official statistics (Global Working Group on Big Data for Official Statistics). The working group, chaired by Statistics Denmark since 2016, has set up a number of teams that have been engaged in the use of selected data sources (satellite data, mobile phone data, scanner data and social media data) as input for the production of official statistics. In addition, since 2014, the working group has organised an annual conference on Big Data for Official Statistics with the participation of statistical institutions as well as private technology enterprises and has established a base of examples of the use of Big Data for official statistics[3].

For many years, Eurostat has also made efforts to develop the use of Big Data for official statistics, e.g. by setting up a steering group in this field, of which Denmark is a member. For the statistical institutions, the most concrete part of Eurostat's engagement in Big Data is the establishment of a so-called ESSnet, which is a number of countries (including Denmark) working together to test the possibilities of using concrete Big Data sources as input in the production of official statistics. Moreover, through the steering group, Eurostat has initiated efforts to test the possibilities at the European level of a legal basis for access to Big Data for the production of official statistics. Finally, Eurostat is also working with competence development in this field, e.g. by offering various data tool courses that are useful for organising and analysing Big Data.

The efforts of the UN as well as the EU facilitate the opportunities for the statistical institutions of the individual countries to gain experience between them and develop employee competences in using Big Data for the production of official statistics.

Finally, a few statistical institutions are particularly advanced in their Big Data activities. In Europe, the Netherlands and the UK and in part Estonia stand out, and outside Europe, Australia and Canada are at an advanced stage. The chance to learn from their example and to work with them must be included in Statistics Denmark's efforts to develop the use of Big Data for official statistics.

[3] You will find further information about the group's work at https://unstats.un.org/bigdata/.

Activities

- Create and maintain an overview of the employees' participation in international activities in relation to Big Data

- Participate in concrete projects in the second round of Eurostat's ESSnet on Big Data for official statistics

- Participate in the relevant task forces in the UN's working group on Big Data for official statistics

- Gain concrete insight into the work with Big Data for official statistics in those countries that are most advanced in this area, e.g. via contacts in the UN's working group on Big Data

- Organise a conference on Big Data, Data Science and statistics with international participation focusing on a Danish audience as a follow-up on the conference held by Statistics Denmark in the autumn 2016 in collaboration with University of Copenhagen and the Confederation of Danish Industry.

Organisation

In line with the work with other data sources, the work with Big Data as a source for official statistics is cross-functional and involves the statistical sections as well as the IT department. In addition, there is a need for management initiatives regarding the efforts to gain access to Big Data (the legal aspects and partnerships) and sources of finance for the combined Big Data efforts in Statistics Denmark as well as the coordination between the international initiatives in Eurostat and the UN and the specific work in Statistics Denmark.

In other words, there are several aspects concerning organisation and financing that need to be clarified and decided.

With the support of the Portfolio secretariat, a project was set up in August 2018, which:

- Makes an analysis of current and future use of Big Data sources for existing and new official statistics, including a conceptualisation of various forms of application of Big Data sources – e.g. in replacement of existing data sources, as a supplement to existing data sources, as a basis for early estimates or new statistics, and an outline of a framework to describe the quality of Big Data sources

- Prepares a presentation which shows potential partners and the potential of Statistics Denmark as a Big Data partner – and against this background

- Carries out a session of interviews (tech lunches) with selected Big Data suppliers, academia and tech corporations as part of the planning of Statistics Denmark's work programme for 2019

It also makes concrete proposals for follow-up on the following activities in the work programme for 2019:

- Organisation of competence partnerships across the organisation and externally (digital task force) to develop and test ideas in strategically selected areas of action

- Preparation of a note based on international experience about conditions involving the legal aspects of the access by official statistics to use Big Data for the production of official statistics

- Active participation in the debate with discussions and events showing good examples of the socio-economic potential of securing statistical and research institutions' access to Big Data

- Organisation of a conference on Big Data, Data Science and statistics with international participation focusing on a Danish audience as a follow-up on the conference in 2016

STATISTICS DENMARK

Statistics Denmark
Sejrøgade 11
2100 Copenhagen Ø

Phone +45 39 17 39 17
www.dst.dk
dst@dst.dk