# Towards Big Data methodology:
## a generic Big Data based statistical process

Piet J.H. Daas and Marco J.H. Puts
Statistics Netherlands, Center for Big Data Statistics

## INTRODUCTION

More and more National Statistical Institutes are investigating the potential of using Big Data. This has resulted in a number of statistics created that use Big Data at various stages of production. A total of 13 Big Data based statistics were identified [1]. These examples have been meticulously compared. This has resulted in a general process for using Big Data as the main data source in official statistics production. This process is composed of two data processing steps, one inference step and one population oriented step.

## GENERIC PROCESS

The basis for the workflow is the one developed for the production of Traffic Intensity statistics based on road sensors [2]. In this process, first a select and reduce step is applied to the raw data (Step 1). This seriously reduces the amount of relevant data remaining; e.g. lots of unneeded data, such as variables and units, are removed. Next information is extracted from the data remaining (Step 2). This, combined with the frame (Step A), is used in the last inference step (Step 3) to produce a statistic. The generic process is shown in Figure 1.

## STEP 2: INFORMATION EXTRACTION

In the next step, the data remaining is processed in such a way that the information required for the specific statistical need is obtained. This may simply involve extracting a value from a dataset, may require applying a filter or a transformation of the data, may involve a model-based approach or a combination of these steps. In the model-based approach, usually an underlying, not-directly measurable variable is estimated. Examples are: i) determine the elementary aggregate of similar products sold, ii) correct for the missing data of road sensors and iii) remove outliers in the location data of ships. These are all very divers methods.

## STEP A: FRAME CREATION

Subsequently, a link has to be made between the population included in the Big data source and the target population of the statistics. Here, two situations may occur. In the first situation, an existing population frame is used to select the data is collected later on in the process (indicated by the dashed line in Fig. 1). The second situation occurs when Big data that is available (i.e. has been collected) is compared with a population frame. This can be challenging and may sometimes even look impossible.

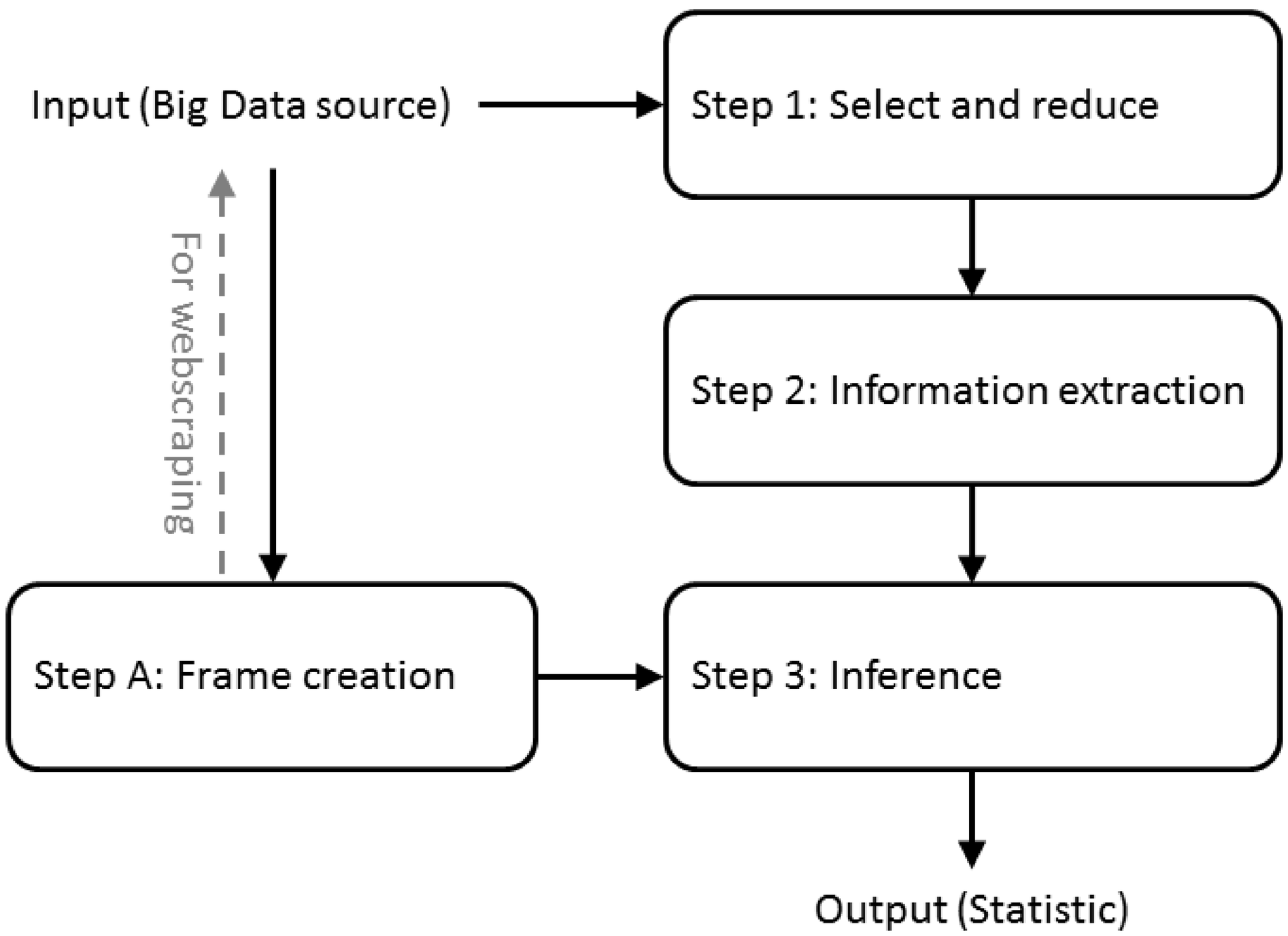


**Figure 1**. Overview of the 4 steps required in Big Data based statistics production

## STEP 1: SELECT and REDUCE

The first step in the process involves a selection at the unit and variable level. As a result the amount of data is often seriously reduced. Goal of this step is either predominantly focused on the removal of unwanted data, the selection of relevant records or a combination of both. It may require the need for specific quality indicators used as selection criteria. Examples of this are: i) selecting products from scanner data, ii) selecting high quality road sensor data and iii) selecting AIS-messages with position data of ships.

## STEP 3: INFERENCE

In the end of the process the frame and the information extracted meet. In this step, nearly always a model-based approach is used in which one aims to infer statistics for the entire population and corrects for any biases. Examples of this step are i) creating a CPI-index based on the products sold, ii) estimating the traffic intensity for road segment, and iii) estimating the number of small and large innovative companies in a country.

## REFERENCES

[1] Daas, P., Puts, M., Maslankowski, J., Salgado, D., Quaresma, S., Tuoto, T., Di Consiglio, L., Brancato, G., Righi, P., Six, M., Kowarik, A. (2020) Report describing the methodological steps of using big data in official statistics with a section on the most important research questions for the future including guidelines. Deliverable K10, Workpackage K, ESSnet Big Data II, 20 November 2020.
[2] Puts, M.J.H., Daas, P.J.H., Tennekes, M., de Blois, C. (2019). Using huge amounts of road sensor data for official statistics. AIMS Mathematics 4, pp. 12-25..