Q 2016
European Conference on Quality in Official Statistics

Quality Control of Web-Scraped and Transaction Data (Scanner Data)

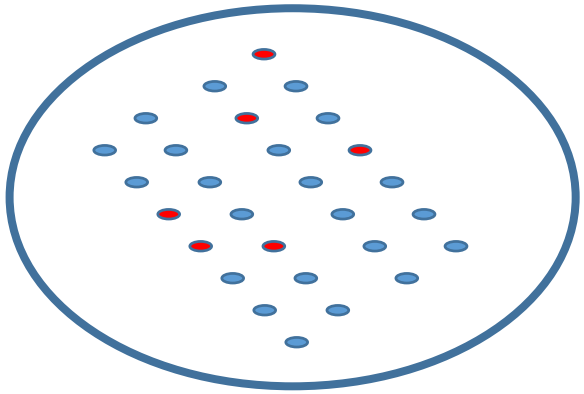
27 – Big Data & Web Scraping

Ingolf Boettcher
Statistics Austria – Consumer Price Statistics
ingolf.boettcher@statistik.gv.at

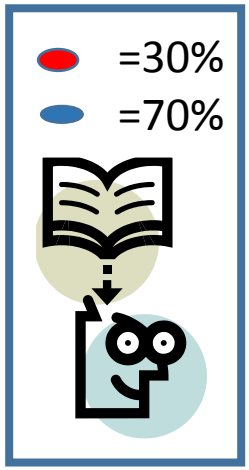
INE Instituto Nacional de Estadística
eurostat
Madrid, May 31 - June 3

Official Statistics production: Where we come from

The universe
(entire statistical population)



Official Statistics



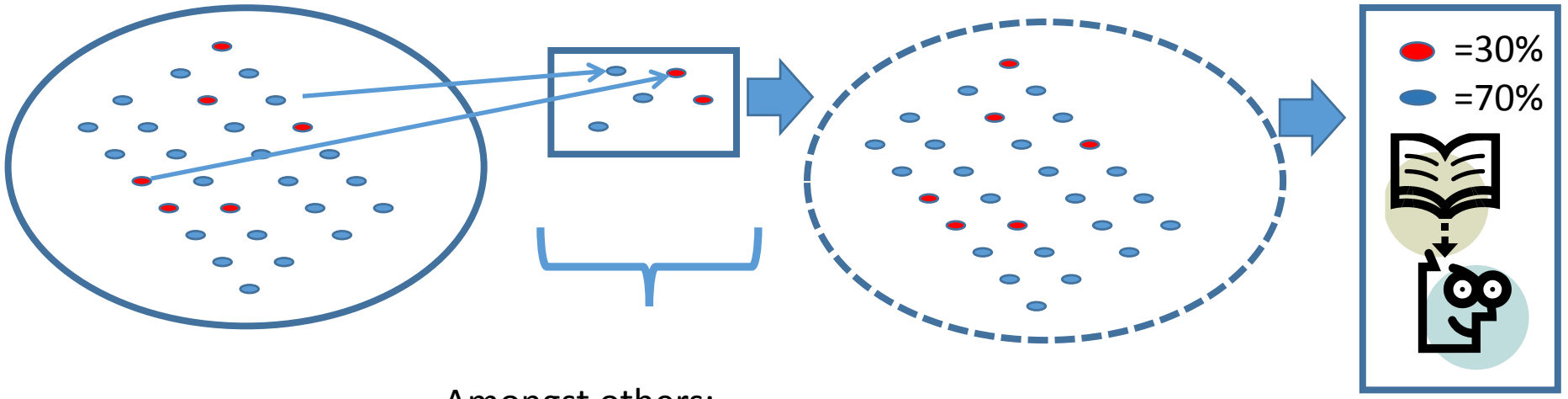
Official Statistics production: Where we come from

The universe
(entire statistical population)

The statistical data
(sample)

The statistical model
(to approximate the universe)

Official Statistics



Amongst others:
Quality control of data input

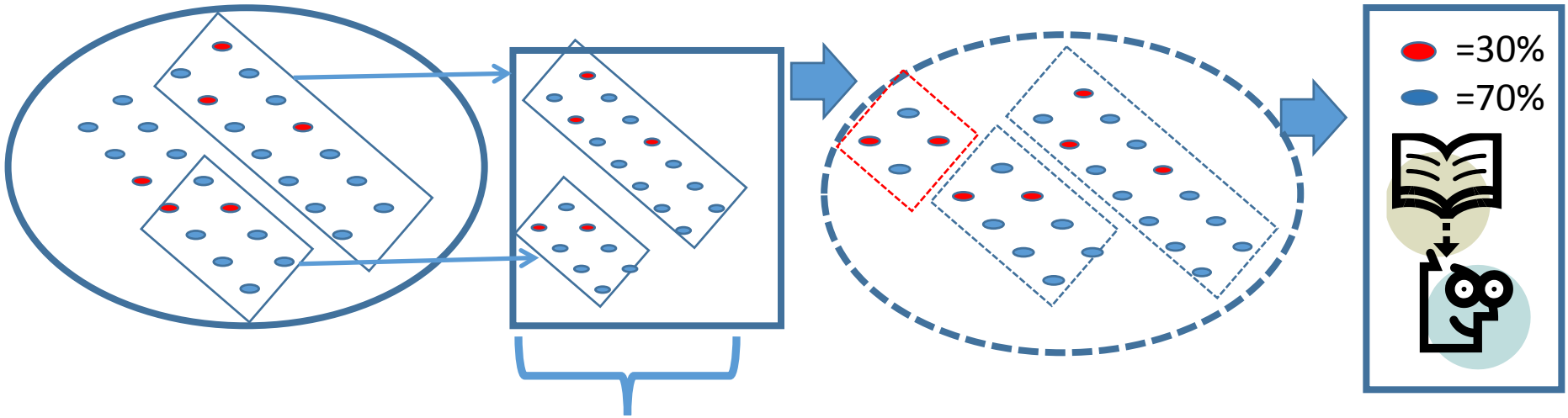
Official Statistics production: with large new data sources

The universe
(entire statistical population)

The statistical data („big data“)

The statistical model
(if necessary….?!)

Official Statistics



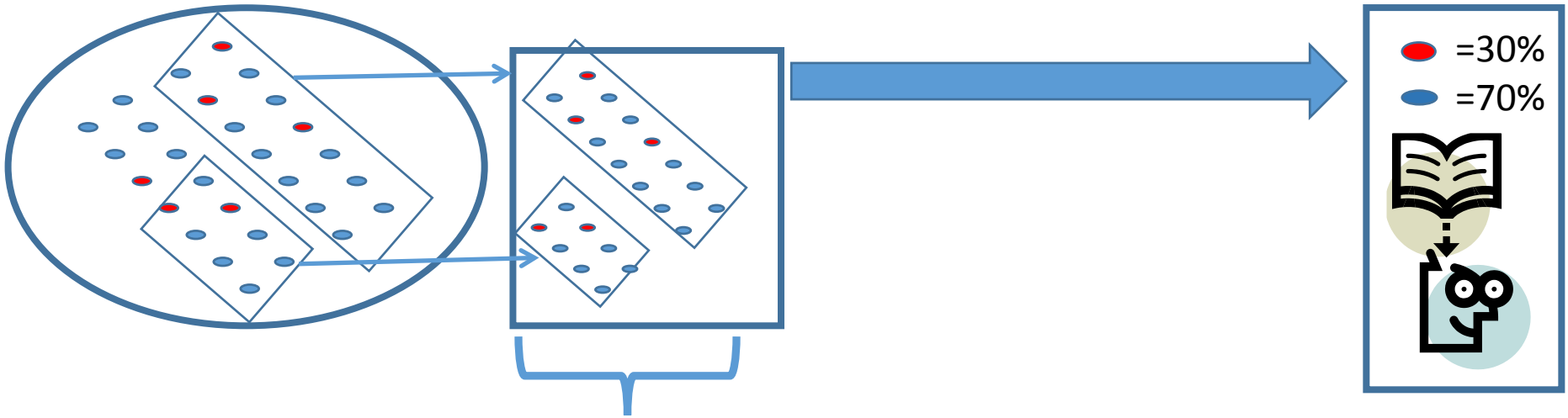
Amongst others:
Quality control of data input

Official Statistics production: with large new data sources – **no need for statistical models? no need for theory?**

The universe
(entire statistical population)

The statistical data („big data“)

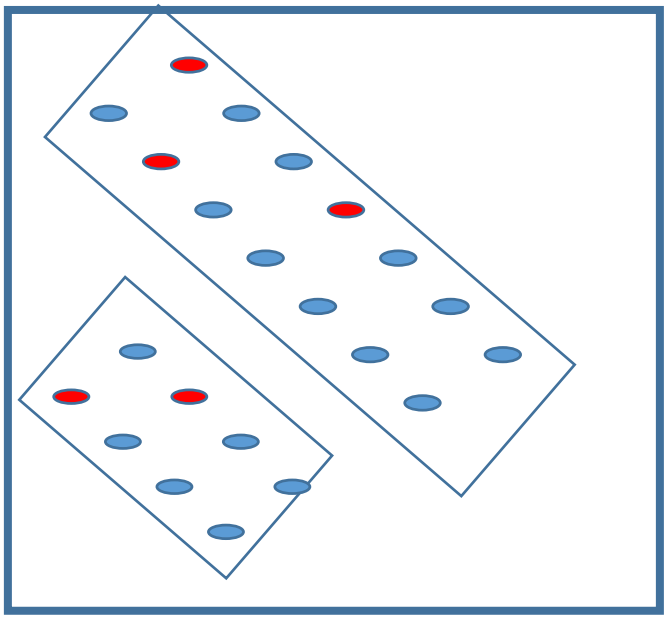
Official Statistics



Amongst others:
Quality control of data input

Quality control of large new data sources

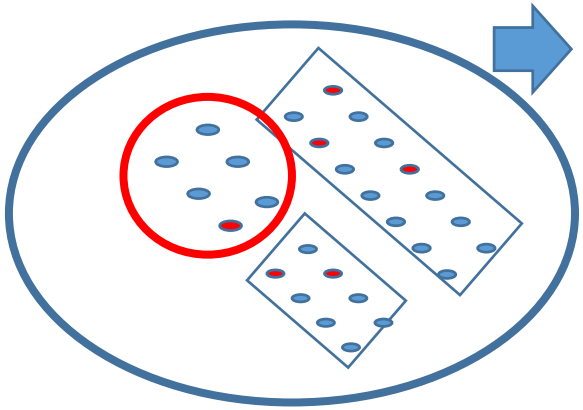
The statistical data
(e.g. supermarket data food and non-food article)



- Is it **relevant?**
- Is it **accurate?**
- Is it **complete?**

Quality control of large new data sources: **relevance**

The statistical data
(e.g. supermarket data food and non-food article)



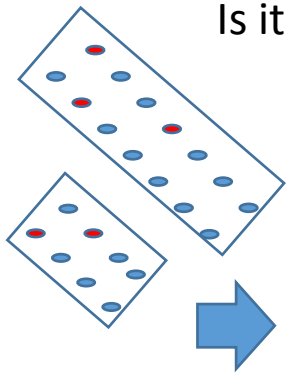
Is it **relevant?**

- Large data-sources do no replace basic methodological work and checks concerning:
  - Coverage bias
  - Measurement error
  - Self selection bias

Large data sources do not make obsolete sound statistical models

Quality control of large new data sources : accuracy/completeness

The statistical data (estimate for Austrian retail market)
(e.g. supermarket scanner data for food and non-food)



Is it **accurate?**

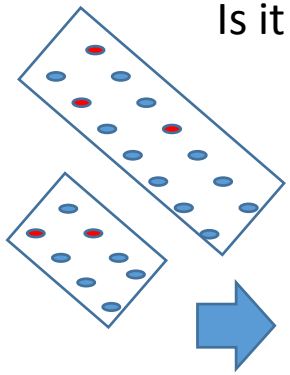
#	Shop ID	Art-Code	Art. retailer classifcation	Product Description	Quantity sold	Sales in EUR
1	212?	1234?	Soft drinks - cola?	Cola, BrandX, 333ML?	123 ?	€129 ?
2	212?	1214?	Soft drinks – cola?	Cola, light, BrandY, L?	255 ?	€126 ?
…	…	…		…	…	…
60.000.000	1234	9965	Bakery products	Brezel, brandZ, 500g	50	€126

60.000.000 data sets every month= 5.000 Articles X 4 Weeks X 1000 Shops X 3 Retailers

Before (with manual price collection):
10.000 data sets = 100 Articles X 1 (monthly collection) X 20 Cities X 5 supermarkets

Quality control of large new data sources : accuracy/completeness

The statistical data
(e.g. supermarket data food and non-food article)



Is it **accurate?**

#	Shop ID	Art-Code	Art. retailer classifcation	Product Description	Quantity sold	Sales in EUR	Accurate & complete?
1	212	1234	Soft drinks - cola	Cola, BrandX, 333ML	123	€129	YES
2	212	1214	Soft drinks – cola	Cola, light, BrandY, L	255	€126	NO

Missing value for „Volume in Liter“

Large new data sources require automation of data cleaning and quality assessment processes

Quality control of large new data sources : accuracy/completeness

Quantitative Approach to Quality control :

1. Define measureable quality dimensions and elements of the data
2. Automate as many consistency and quality checks as possible

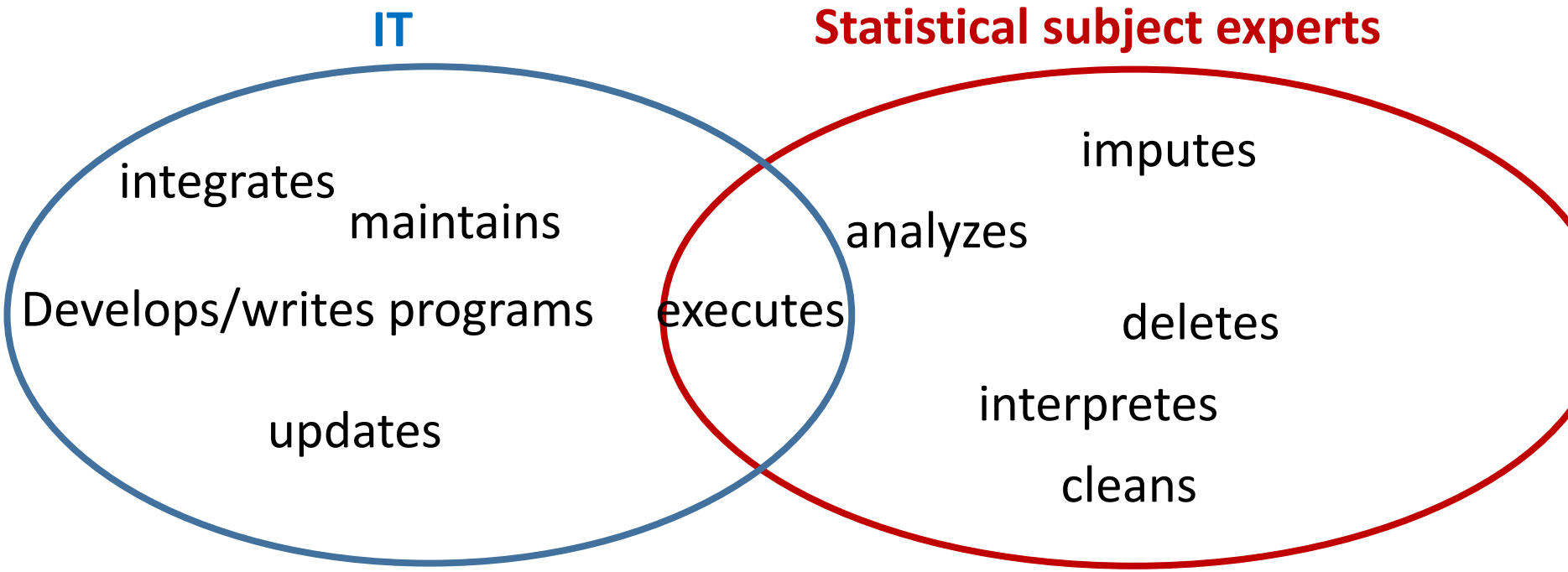
Examples:

-Extent in % of erroneous / inconsistent data is monitored and excluded
-average # of missing values per data set
-unreasonable changes of summary statistics
-Number and level of target values measured against historical values
-% of month to month attrition rates in product groups

3. Ability to adapt automated processes to ever-changing data structures and sources

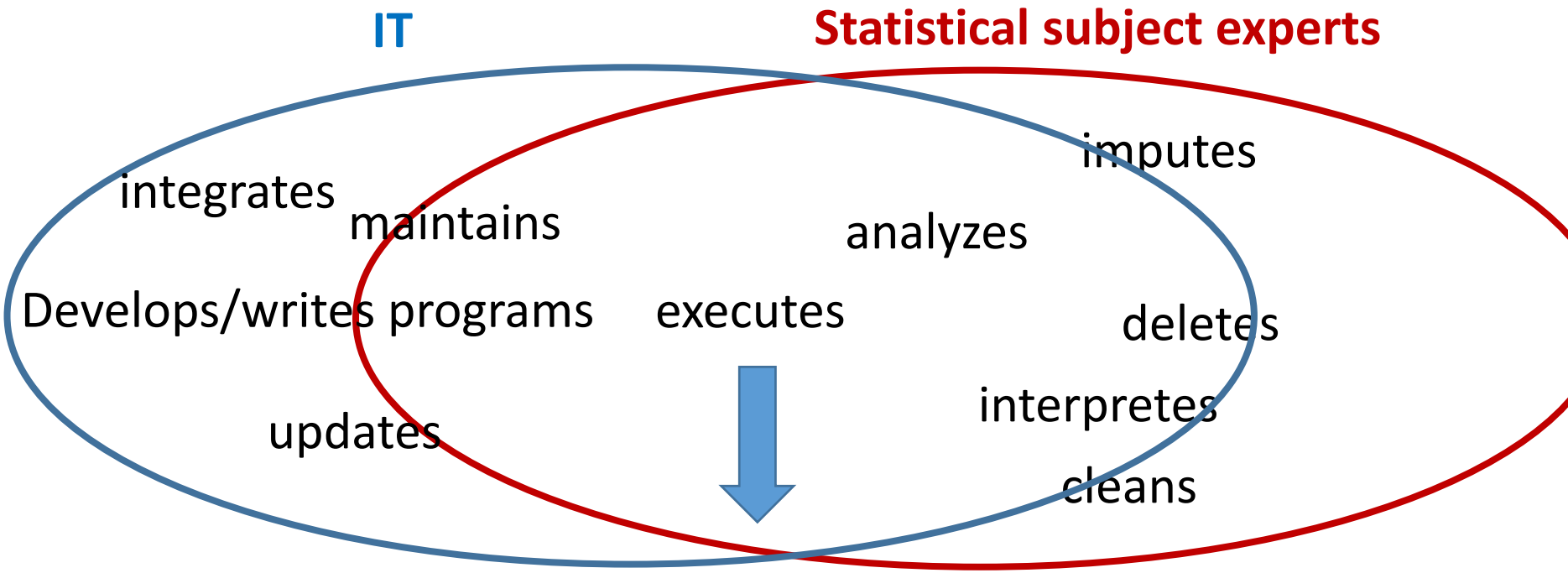
Quality control of large new data sources : accuracy/completeness

3. Adapt automated processes to changing data structures and sources



Quality control of large new data sources : accuracy/completeness

3. Adapt automated processes to changing data structures and sources = Data science



„Data science“ (in official statistics)–>integrate, clean and analyze continuously changing (non-standardized) large data sources and turn them into compliant standardized official statistics

Quality control of large new data sources : accuracy/completeness

3. Adapt automated processes to changing data structures and sources = Data science

Examples	
Scanner data -retailer continuously update data-base structures to own data-warehouse needs -high attrition rate of single articles, shops, product classes	**Web-scraping** -frequently changing web-site architecture and product presentation -high attrition rate of single articles and categories

Q 2016
European Conference on Quality in Official Statistics

Quality Control of Web-Scraped and Transaction Data (Scanner Data)

27 – Big Data & Web Scraping

Ingolf Boettcher
Statistics Austria – Consumer Price Statistics
ingolf.boettcher@statistik.gv.at

INE Instituto Nacional de Estadística
eurostat
Madrid, May 31 - June 3