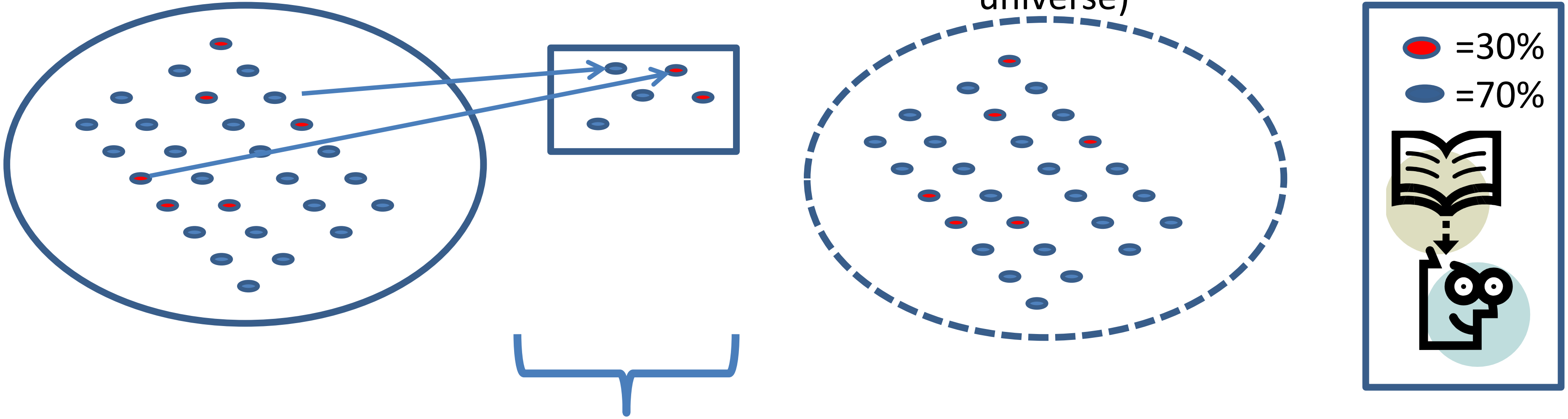
Josef Auer
Ingolf Boettcher

2017 Ottawa Group

From price collection to price data analytics

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

● =30%

● =70%

Amongst others:
Quality control of
data input

Integration of large new data sources

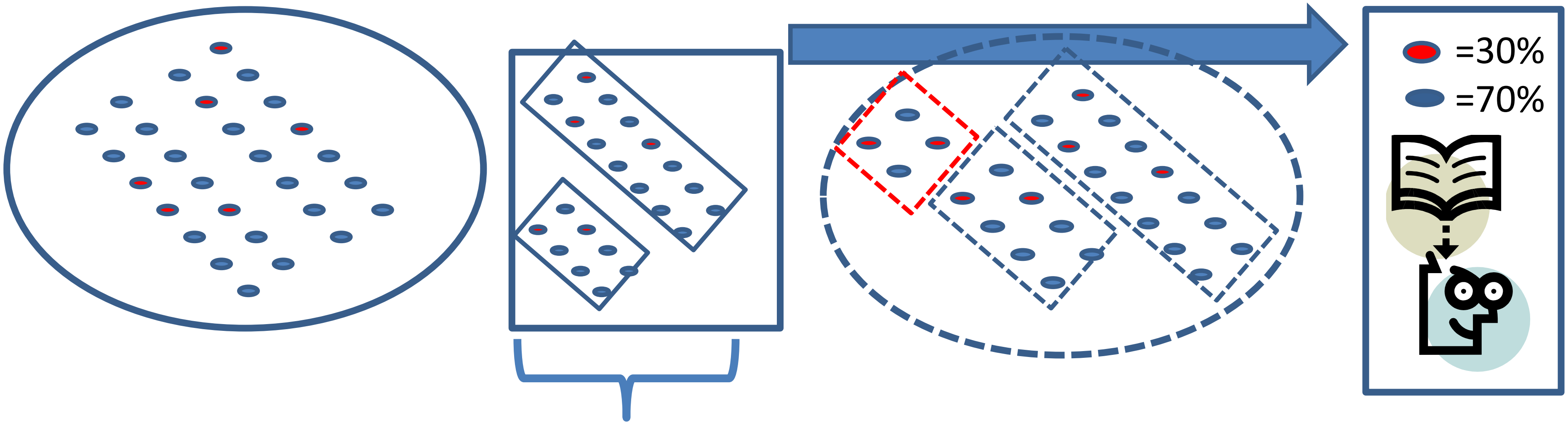
no need for statistical models?
no need for theory?

The universe
(entire statistical population)

The statistical data („big data“)

The statistical model
(if necessary….?!)

Official Statistics



🔴 =30%
🔵 =70%

Amongst others:
Quality control of
data input

Integration of large new data sources

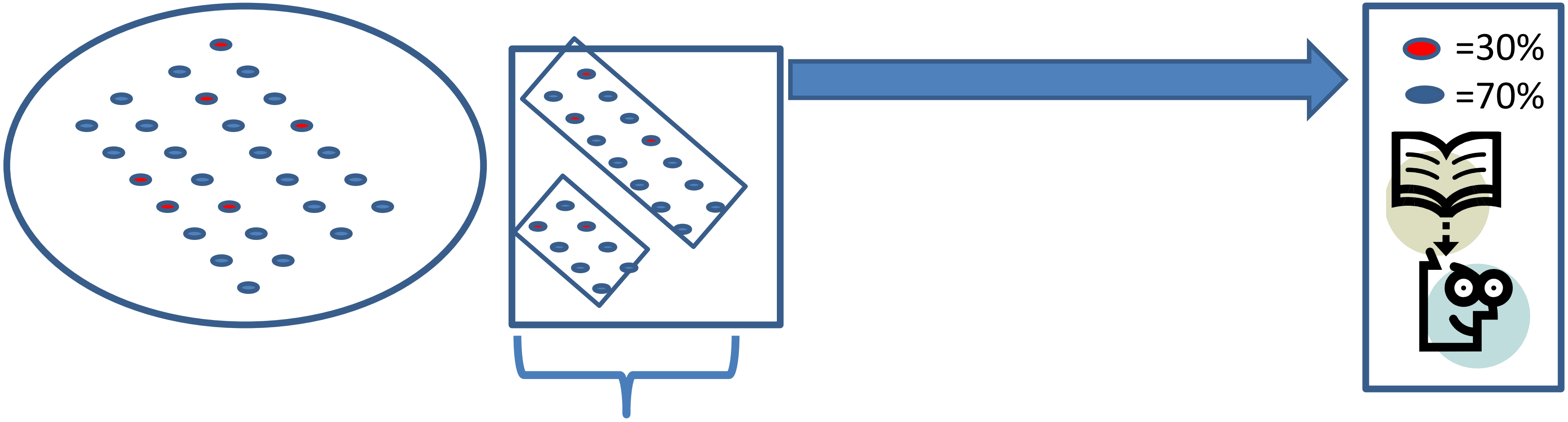
no need for statistical models?

no need for theory?

The universe
(entire statistical population)

The statistical data („big data“)

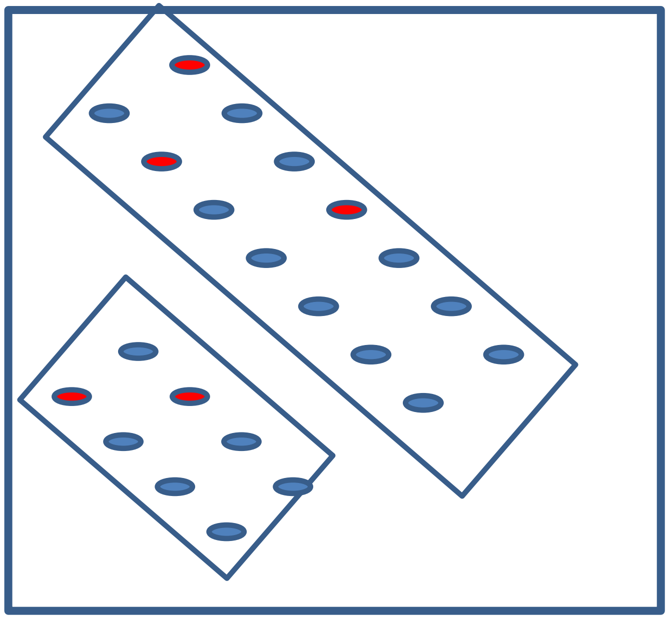
Official Statistics



Amongst others:
Quality control of
data input

Integration of large new data sources

Quality control of scanner data and the web-scraped data
→ new measurment methods necessary



→ Is it **relevant?**

→ Is it **accurate?**

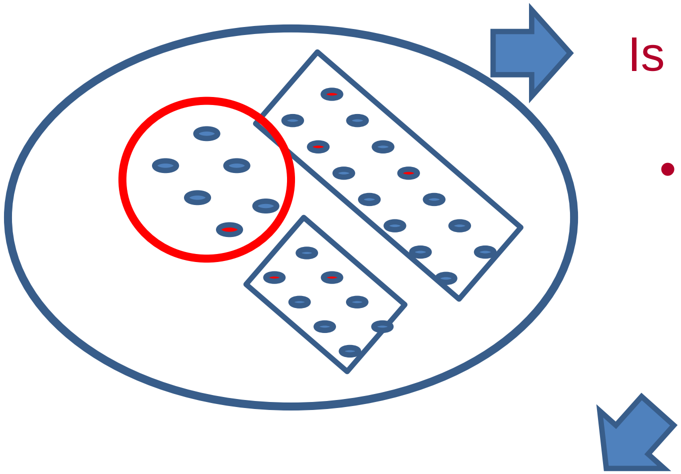
→ Is it **complete?**

Relevance of scanner data

Quality problem – Data Relevance	Measurement Method
Transaction data may contain transactions that are out of scope. -e.g. expenditures for business purposes (out of scope for consumer price indices)	**Information by data providers;** otherwise unresolved

Integration 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

Relevance of web-scraped data

Quality problem – Data Relevance	**Measurement Method**
are products offered online really sold and by whom?	**Information by data providers;** otherwise unresolved

Accuracy of scanner data

Quality problem – Data Accuracy	**Measurement Method**
Volume and variety of data sets are too large to identify and clean erroneous/ untrustworthy/ inconsistent data sets with conventional methods.	Extent in % of erroneous / inconsistent data is monitored and excluded

Accuracy of web-scraped data

Quality problem – Data Accuracy	**Measurement Method**
Website content may be IP-specific (a user who frequently checks a website or a web-scraper might lead to different price displays than first-time users)	Comparison of automatically and manually collected data

Completeness of scanner data

Quality problem – Data Completeness	Measurement Method
Volume and variety of data sets are too large to identify missing values with conventional methods. (Scanner data: natural attrition of Unique identifiers is extremely high)	Number and level of target values are measured against historical values from previous deliveries

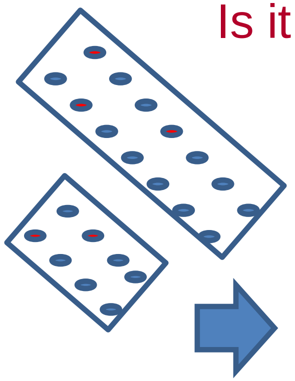
Completeness of web-scraped data

Quality problem – Data Completeness	**Measurement Method**
Websites change frequently Relevant variables and URLs might not be identified and scraped	Number and level of target values are measured against historical values from previous deliveries

Implementation 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

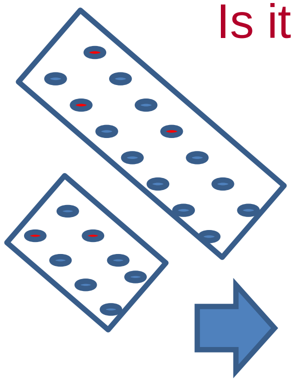
60.000.000 data sets every month= 5.000 Articles X 4 Weeks X 1000 Shops X 3 Retailers

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

Implementation 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

Implementation of large new data sources : **accuracy/completeness**

Analytical approach to quality control

1.Define measureable quality dimensions and elements of the data

2.Automate as many consistency and quality checks as possible

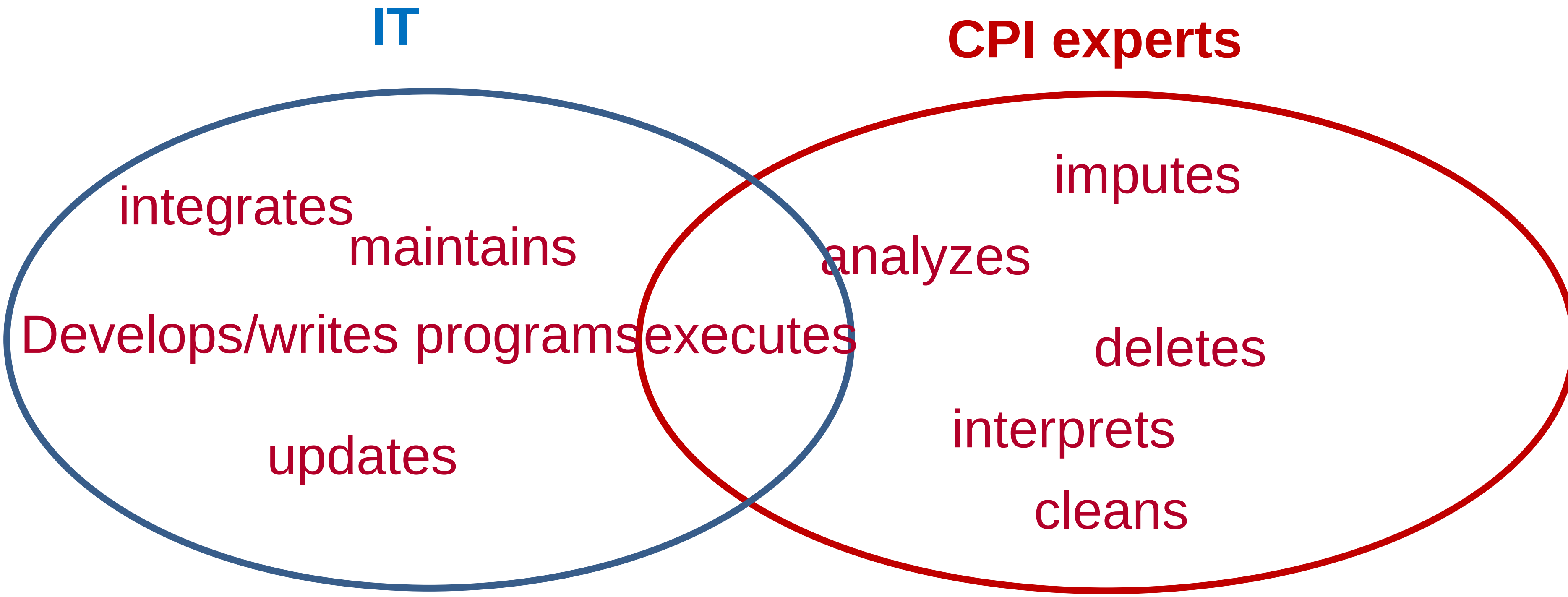
Examples:

-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

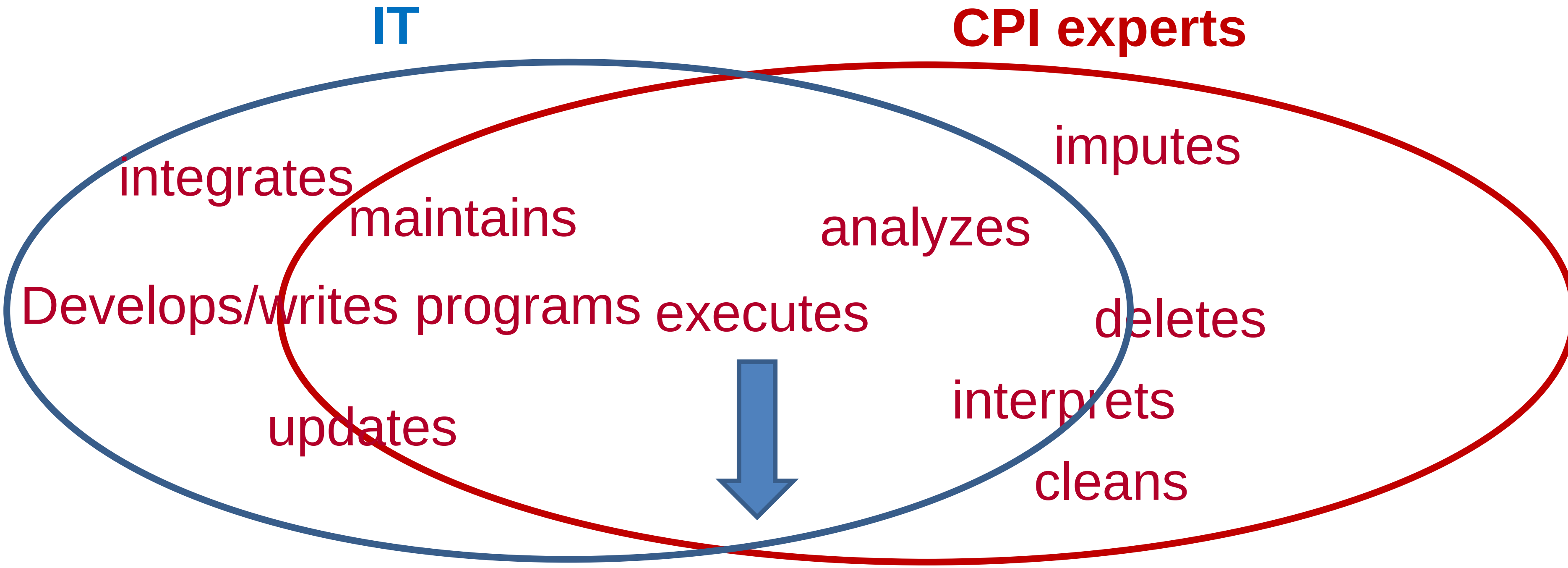
Implementation of large new data sources : **accuracy/completeness**

3. Adapt automated processes to changing data structures and sources



Implementation of large new data sources : **accuracy/completeness**

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



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

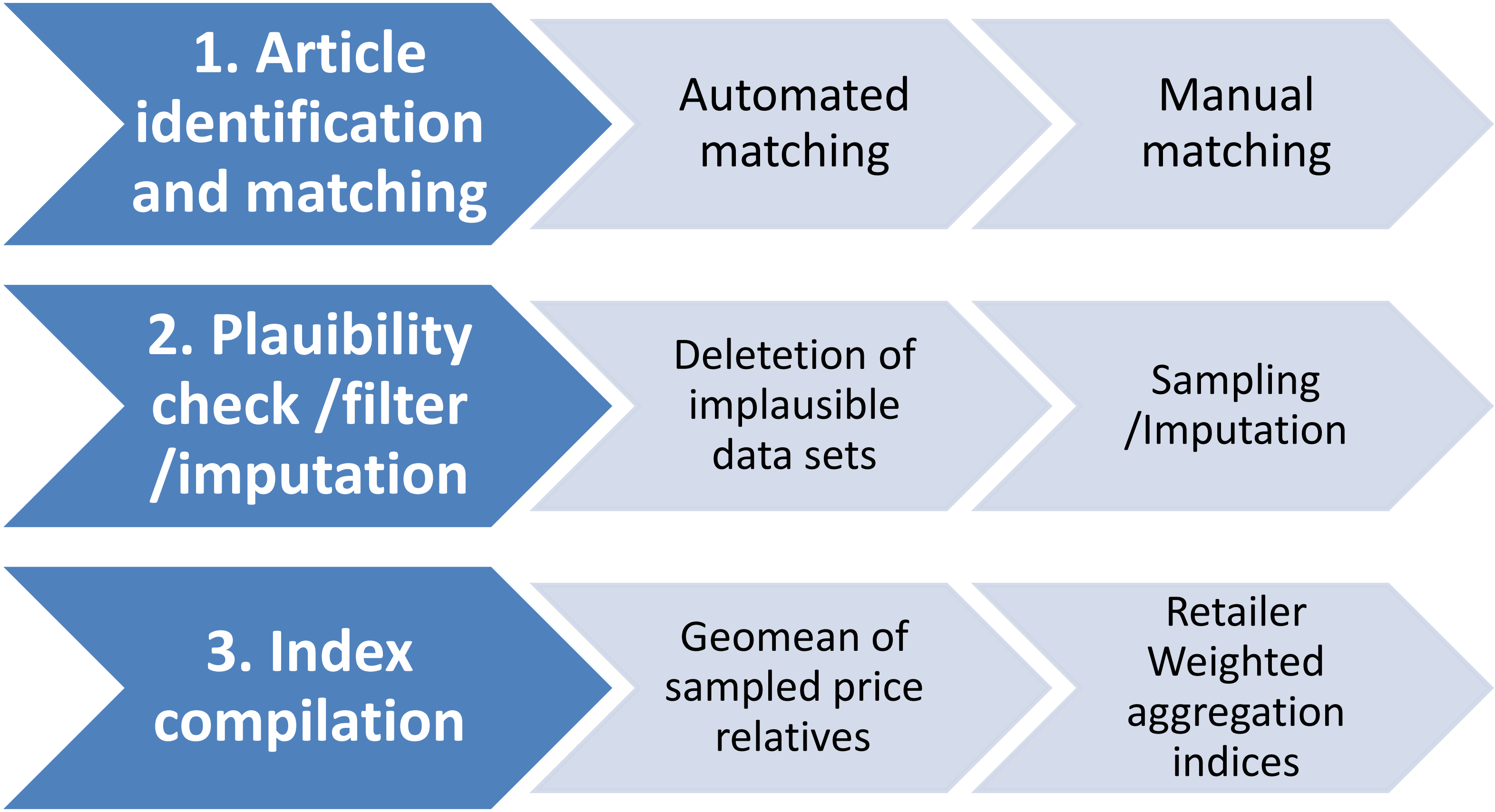
Implementation of large new data sources :

3. Adapt automated price index compilation 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

Price index compilation with scanner data

new working steps



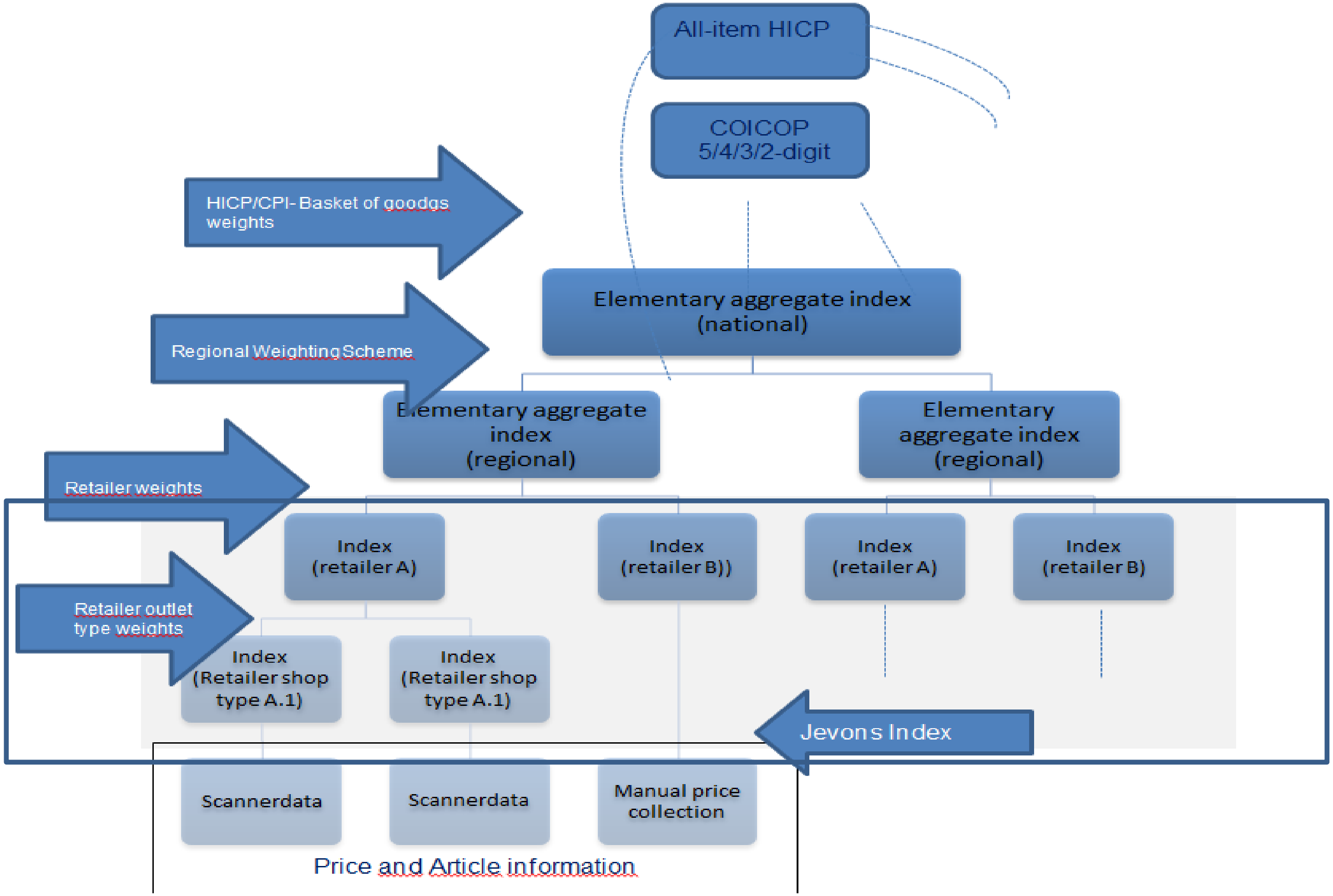
1. Article identification and matching → Automated matching → Manual matching

2. Plauibility check /filter /imputation → Deletetion of implausible data sets → Sampling /Imputation

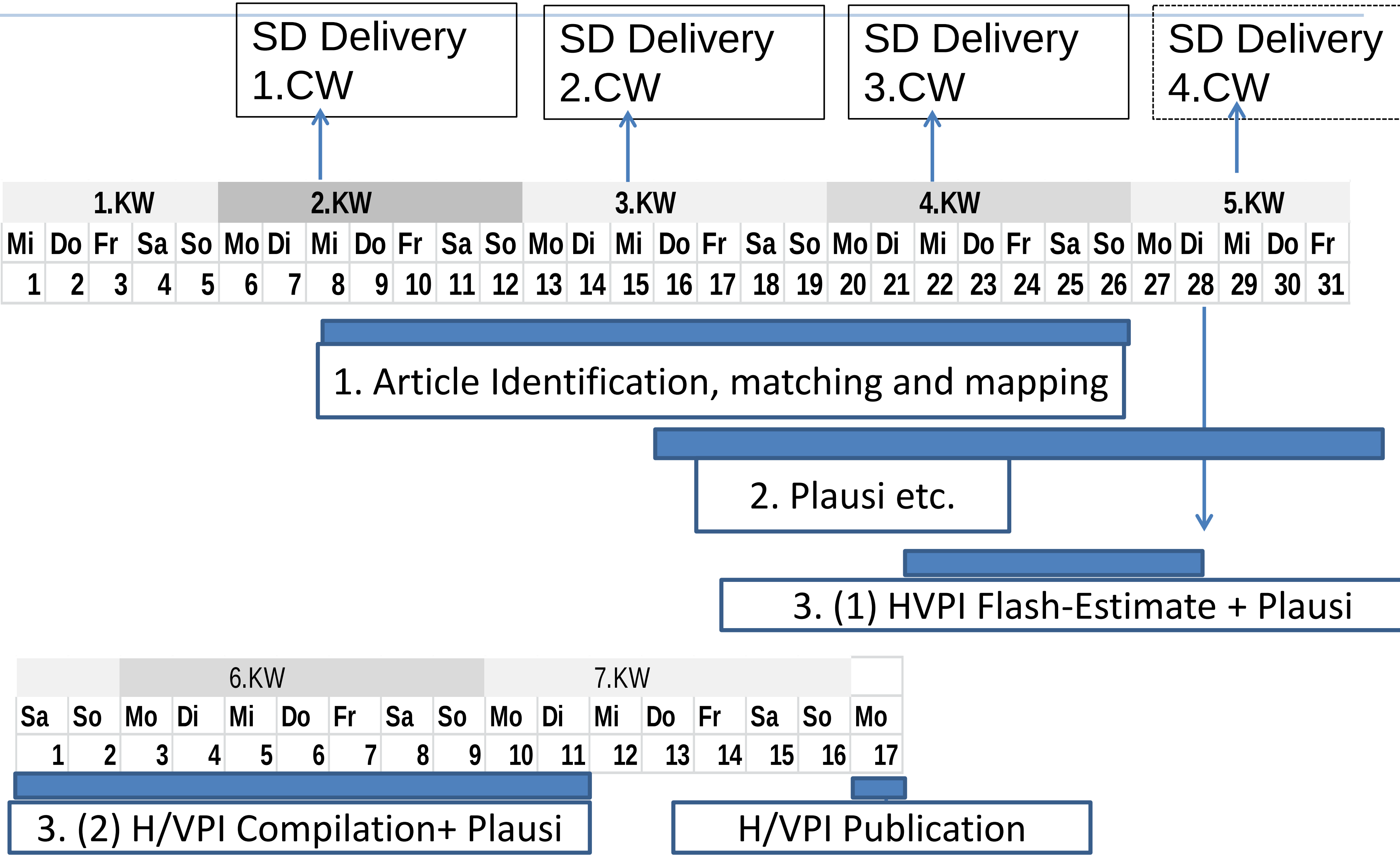
3. Index compilation → Geomean of sampled price relatives → Retailer Weighted aggregation indices

Price index compliation with scanner data

new strata


All-item HICP
COICOP 5/4/3/2-digit
HICP/CPI- Basket of goodgs weights
Elementary aggregate index (national)
Regional Weighting Scheme
Elementary aggregate index (regional)
Elementary aggregate index (regional)
Retailer weights
Index (retailer A)
Index (retailer B))
Index (retailer A)
Index (retailer B)
Retailer outlet type weights
Index (Retailer shop type A.1)
Index (Retailer shop type A.1)
Jevons Index
Scannerdata
Scannerdata
Manual price collection
Price and Article information

Price index compliation with scanner data



SD Delivery 1.CW

SD Delivery 2.CW

SD Delivery 3.CW

SD Delivery 4.CW

1.KW					2.KW							3.KW							4.KW							5.KW				
Mi	Do	Fr	Sa	So	Mo	Di	Mi	Do	Fr	Sa	So	Mo	Di	Mi	Do	Fr	Sa	So	Mo	Di	Mi	Do	Fr	Sa	So	Mo	Di	Mi	Do	Fr
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

1. Article Identification, matching and mapping

2. Plausi etc.

3. (1) HVPI Flash-Estimate + Plausi

		6.KW							7.KW							
Sa	So	Mo	Di	Mi	Do	Fr	Sa	So	Mo	Di	Mi	Do	Fr	Sa	So	Mo
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

3. (2) H/VPI Compilation+ Plausi

H/VPI Publication

Contact:

Josef Auer
josef.auer@statistik.gv.at

Ingolf Boettcher
ingolf.boettcher@statistik.gv.at

From price collection to price data analytics

t