Experiences in the Use of Big Data for Official Statistics

Antonino Virgillito

Istat

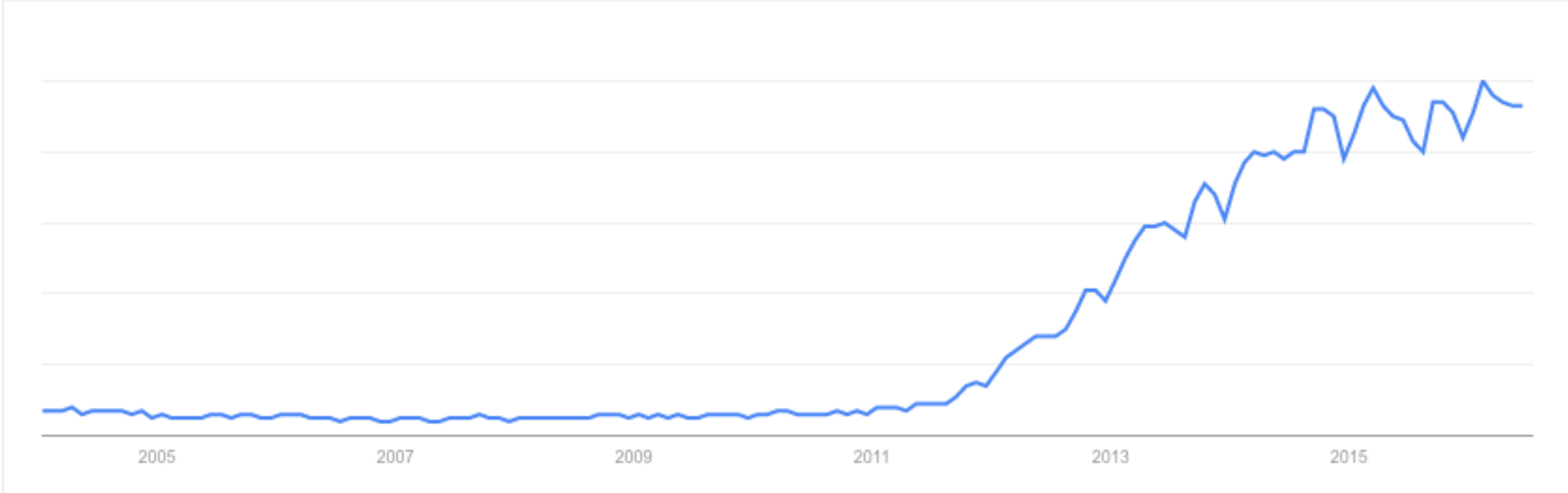
Introduction

The use of Big Data sources for the production of official statistics has been the subject of a lively discussion within the statistical community in recent years, producing a significant body of study and work

Particularly, Istat has developed an extensive experience in this area, with several ongoing initiatives

We present some of these experiences and highlight results and lessons learned

The Big Data Global Trend



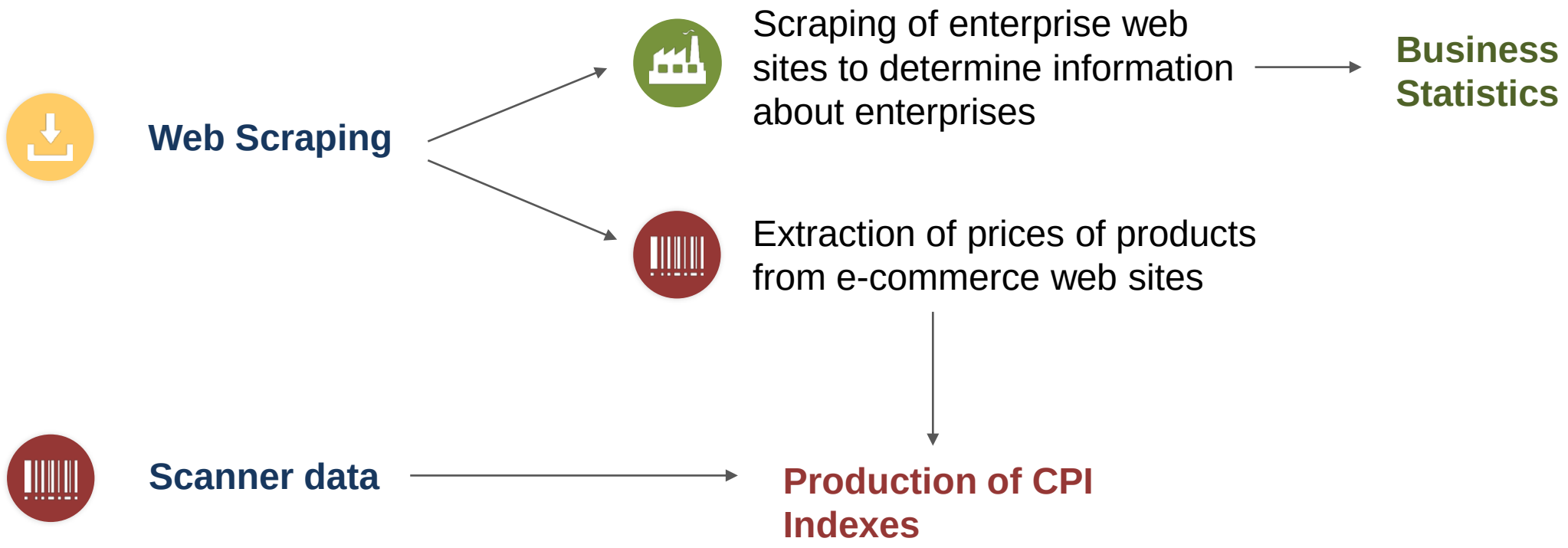
Result of searching “Big Data” in Google Trends

The Path to Big Data in Official Statistics

UNECE Big Data Project
(2014 - 2015)
Demonstrate feasibility of production based on Big Data sources

ESSnet Big Data project
(2016 - 2018)
Integration of Big Data in the regular production of official statistics

Experiences in Istat



Web Scraping

Two approaches

Scrape textual content of a large number of web sites and analyze it offline to determine some information through text mining techniques

Extract specific information from semi-structured web sites through custom software or automation tools (robots)

Web Scraping Enterprise Web Sites

General Objective:

to investigate whether web scraping, text mining and inference techniques can be used to collect, process and improve general information about enterprises

Enterprises
Web sites

Business Statistics Surveys

National Business Register

Crawling → Scraping → Indexing → Searching

Using Scraped Data

Use case 1: URLs Inventory

Use case 2: Web sales - ECommerce

Use Case 3: Social Media Presence

Use Case 4: Job Advertisement

Use Case 1: URLs Inventory

- Main Identified Population (ICT Survey):
  - Enterprises with at least 10 Employees
  - Not all of them have a web site, but for those of them that do have it, the URLs of the web sites are not fully available
- The URLs Retrieval problem:
  - Given a set of identifiers (denomination, fiscal code, economic activity, etc.) for the enterprise X, searching the Web for
    - Retrieving a set of associated URLs
    - Estimate (if any) which is the URL corresponding to the web site of X

URLs Inventory: Technique and Results

- Steps
  - Query a search engine for enterprise name
  - Crawl the returned pages and score them according to content
  - Classify the results with machine learning approach
- Machine learning step
  - Logisitic model fitted on a training set, and then applied to the set of all other enterprises
  - Application of the model to the set of unlabeled (i.e. not belonging to the training set) enterprises
- Total number of identified URLs equal to about 105,000 out of 130,000 websites pertaining to the enterprises population
  - 81% coverage

Use Case 2: Web sales - Ecommerce

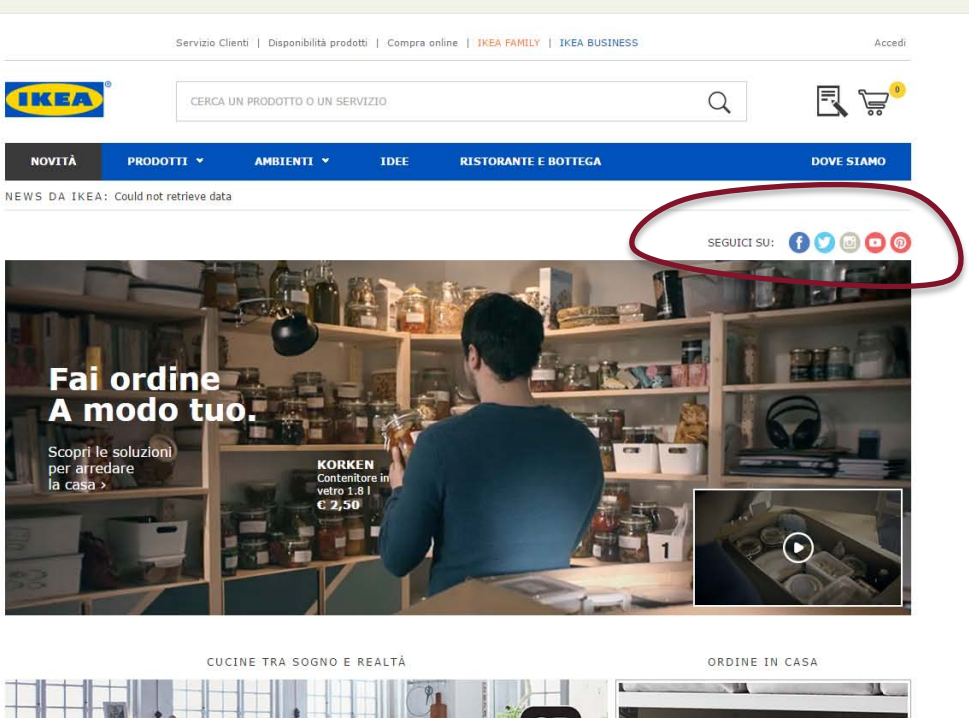
Predict whether an enterprise provides or not web sales facilities on its website

	Use of a Website or Home Page		
B7.	**In January 2013, did your enterprise have a Website or Home Page?** (Filter question)	Yes	No -> go to B9
B8.	**In January 2013, did the Website or Home Page have any of the following?**	Yes	No
	*[6] a) Online ordering or reservation or booking, e.g. shopping cart		
	b) A privacy policy statement, a privacy seal or certification related to website safety		
	c) Product catalogues or price lists		
	d) Order tracking available on line		
	e) Possibility for visitors to customise or design the products		
	f) Personalised content in the website for regular/repeated visitors		
	g) Advertisement of open job positions or online job application *- Optional*		

ICT Survey

Use Case 3: Social Media Presence

Information on existence of the particular enterprises in social media (mainly Twitter and Facebook)



Use Case 4: Job Advertisement

Investigating how enterprises
use their websites to handle the job advertisements, and in particular if they publish job advertisement or not

Technique and Results

- Prediction realized through different classification algorithms
  - logistic model, classification trees, random forests, boosting, bagging, neural net, Naive Bayes, SVM
- Evaluation of algorithms performance according to different indicators
- Quality of results still to be improved
  - Social media use case almost ready for production

Web Scraping Prices

- The collection of data from the Internet through the extraction of structured content from web pages is an established technique for statistical data collection
  - Replace repetitive centralized tasks
  - Possibility for getting more data
- Price data is particularly attractive…
  - A lot of prices on the Internet!
  - Common practice in European NSIs

Web Scraping Prices at Istat

17 types of products currently collected through scraping in production

Consumer electronic products: collection of prices from 4 different e-commerce web sites, including Amazon.

Energy sector: collection of gas tariffs from the conditions published on the web site of one major Italian energy provider.

Fiscal sector: current taxes rates collected from the Minister of Finance web site.

Financial sector: cost of bank accounts, collected from a web portal that allows to compare offering and costs of bank accounts for consumers.

Transport sector: cost of tickets for trains and flights (Experimental)

Problems

- Sustainability
  - The more we develop scraping solutions, the more maintenance is required
  - Maintenance requires dedicated IT resources
- Scale
  - Scraping for prices is substantially a replacement for manual collection activity
  - Difficult to collect large data size
  - Data must be selected manually before collection

Web Scraping Prices
Results and Next Steps

- Significant improvements in efficiency have been achieved so far through tools and techniques that are now mature and familiar
- The risk is that we are not able to reach to the next level of scale and exploit the full potential of web data
- Are we ready to try new approaches?

Scanner Data

Starting from 2014 Istat is collecting data from the six major chains of retail data distribution

Weekly sales and turnover at individual product level

Production in parallel with traditional method is going to start during 2017

Scanner Data

Data size

1,4 Billion records of raw data currently in the database

Will double during 2017 for the acquisition of the full sample

Will continuously grow along years

200Gb after processing

Global historical data warehouse of price data

Not so big in absolute terms, but still uncommon for an NSI

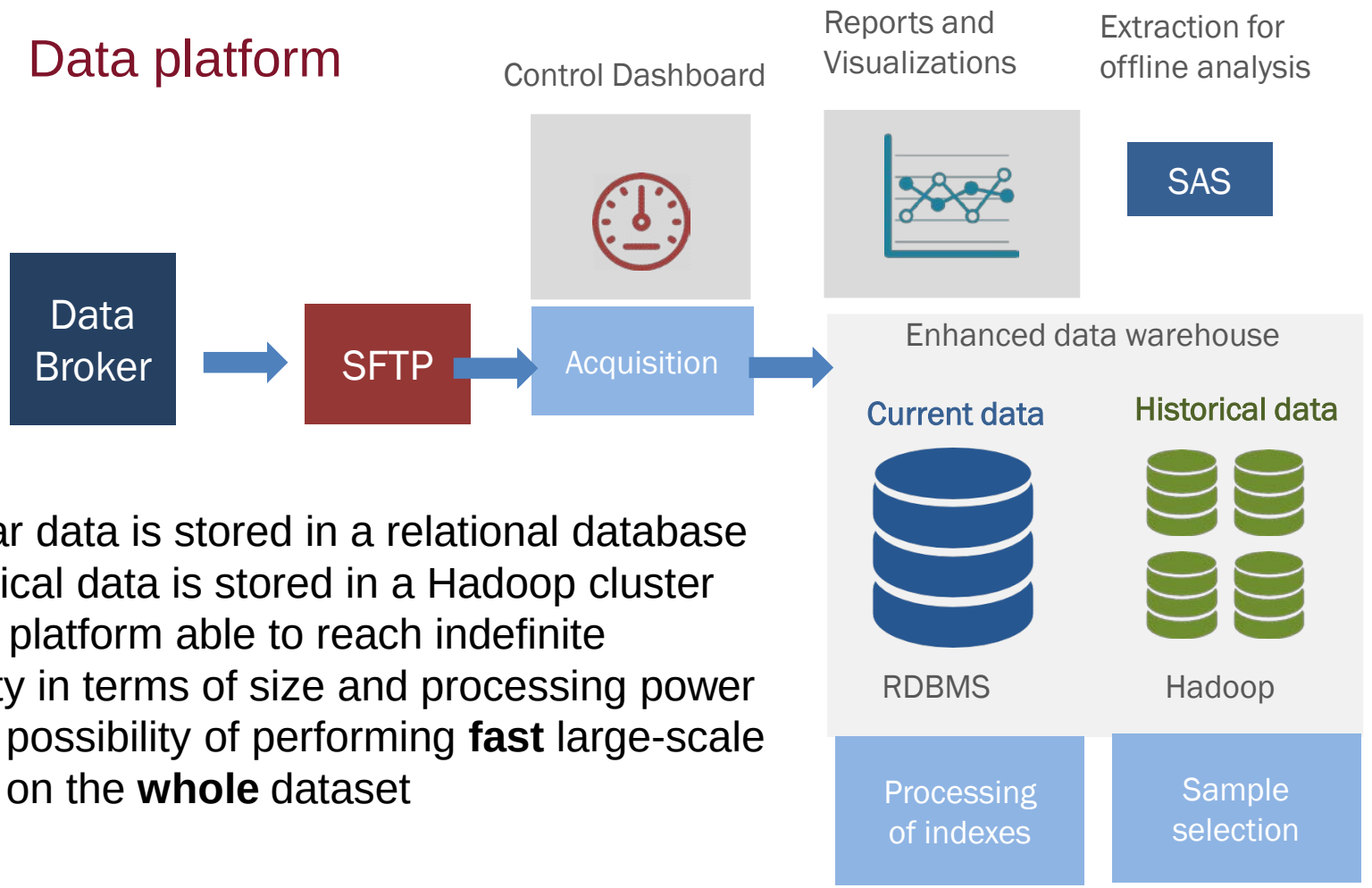
Can be used also in other surveys

Scanner Data

The problem with size...

- Query time is not satisfactory
  - IT support required
  - Aggregated intermediate results in tables
  - More space!
- Difficult to extract data for analysis
- Data growth is not predictable
- Difficult maintenance for DBAs

IT Architecture



Current year data is stored in a relational database while historical data is stored in a Hadoop cluster

- Big data platform able to reach indefinite scalability in terms of size and processing power
- Enables possibility of performing **fast** large-scale analysis on the **whole** dataset

Scanner Data: Results So Far

- Big Data technology is being exploited not only for storing data but also to improve the efficiency of data processing
  - Test of different sampling schemes has been made since now by methodologist using **desktop SAS** on a small portion of the data set, taking **several hours** for its execution
  - We are implementing procedures in **Spark** to compute quality indicators on the whole data set that will allow to compare the performance of the different sampling strategies
- The IT architecture set up for this project is now part of Istat production architecture
  - Training of administrators to support the infrastructure

Other Experimental Topics

Istat is carrying out other activities on Big Data, still in experimental stage

- Use of **mobile phone data** for statistics on population and tourism
- Soil consumption statistics based on analysis of **satellite images**
- Prediction of economic indicators based on **Google Trends**
- **Scraping** web sites of agritourisms for tourism statistics
- Statistics on road traffic based on analysis of **traffic cameras**
- Energy consumption statistics based on **smart meters**
- Estimation of consumer confidence from sentiment analysis of **Twitter data**

Lessons Learned

Accessing data is **always** difficult

- Relevant data is normally is subject to privacy constraints
- Datasets may be big and difficult to move and to treat
- Production statistics requires continuity, often difficult to achieve
- Internet data is only apparently “easy” to get

Lessons Learned

Different notions of quality could be considered

- Impossible to get the same level of control we are used to, on sources like Internet and social media data
- Planned publication of experimental statistics

Lessons Learned

Technology requires solid governance

- IT platforms are complex to install and manage
- Cloud solutions are not an option for privacy-critical data
- Internal skills are needed

Lessons Learned

Focused capacity building strategy is crucial

- Non familiar paradigms: statisticians are in general more confortable with desktop computing tools
- Wide range of skills are needed and it is difficult to devise a coherent strategy

Conclusions

- The statistical community has made major steps forward in the development of processes based on Big Data sources
- The work in recent years was mainly useful to separate the myths generated by high initial expectations from the really useful and innovative aspects
- High-quality results require significant effort while stakeholders expect fast output
- The real challenge for NSIs in the near future is to find a sustainable way to achieve results quickly while building a solid road for the future