Application of Big Data in Official Statistics

14th Meeting of the Heads of National Statistical Office of BRICS Countries

Eduardo L.G. Rios-Neto

President of IBGE

October 11th, 2022

SUMMARY

1. BIG DATA – Definition
2. ADMINISTRATIVE DATA IBGE
3. GEOSPATIAL STATISTICAL FRAMEWORK (GSF)
4. OTHER BIG DATA ANALYSIS – “Webscrapping”
5. REGIONAL HUB FOR BIG DATA IN BRAZIL

1- BIG DATA - DEFINITION

Social Science Research 59 (2016) 1–12

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

The role of administrative data in the big data revolution in social science research

Roxanne Connelly [a, *], Christopher J. Playford [b], Vernon Gayle [c], Chris Dibben [d]

[a] Department of Sociology, University of Warwick, Social Sciences Building, The University of Warwick, Coventry, CV4 7AL, UK
[b] Administrative Data Research Centre – Scotland, University of Edinburgh, 9 Edinburgh Bioquarter, Little France Road, Edinburgh, EH16 4UX, UK
[c] School of Social and Political Science, University of Edinburgh, 18 Buccleuch Place, Edinburgh, EH8 9LN, UK
[d] School of Geosciences, University of Edinburgh, Geography Building, Drummond Street, Edinburgh, EH8 9XP, UK

Made Data Experimental	**Made Data** Observational (e.g. Social Surveys)	**Found Data** Administrative Data	**Found Data** Other Types of Big Data
• Data are collected to investigate a fixed hypothesis. • Usually relatively small in size. • Usually relatively uncomplex. • Highly systematic. • Known sample / population.	• Data may be used to address multiple research questions. • Data may be very large and complex (but usually smaller than big data). • Highly systematic. • Known sample / population.	• Data are not collected for research purposes. • May be large and complex. • Semi-systematic. • May be messy (i.e. may involve extensive data management to clean and organise the data). • Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data inkage). • Usually a known sample / population.	• Data are not collected for research purposes. • May be very large and very complex. • Some sources will be very unsystematic (e.g. data from social media posts). • Very messy / chaotic. • Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage). • Sample / population usually unknown.

IBGE & NSO

Fig. 1. Characteristics of quantitative social science data resources.

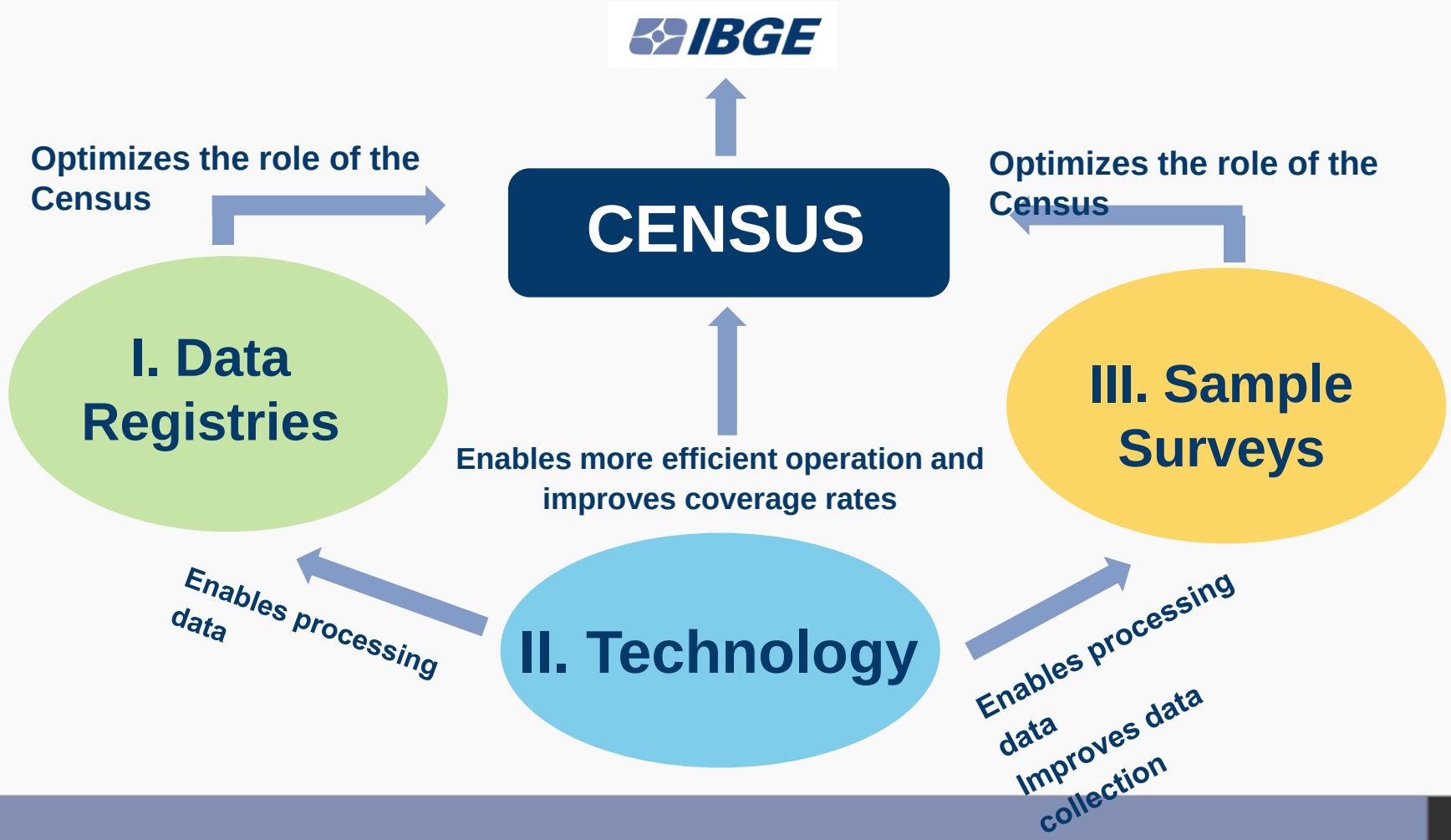
4.2. *Administrative social science data*

As we have described, administrative social science data differs from made data as it is not originally collected for the purpose of research (see panel 3 of Fig. 1). Researchers generally have no input into the design, structure and content of administrative social science data. These datasets can be large, however they are often not as large as the types of big data collected through for example, social media, GPS tracking, or supermarket transactions. These data are also likely to be more complex than the well curated social survey data resources which researchers may be accustomed to. The data may be messy and the use of administrative social science data resources is likely to involve substantial data management (or enabling) to clean and organise the data into the format required for analysis.

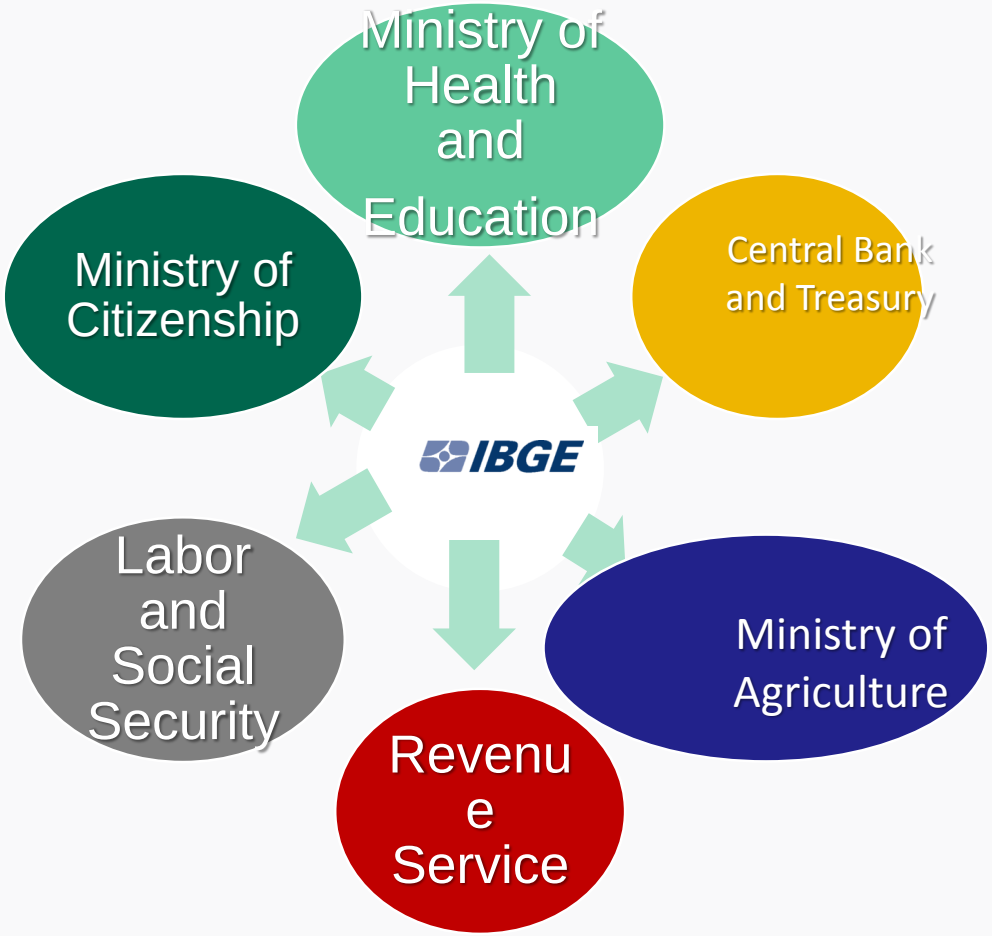
In many ways one of the key benefits of administrative social science data is that it will be complementary to sources of made data (e.g. social survey data). Social survey data can provide the means to collect detailed information not available in administrative data. Administrative data can provide independent measures and additional information (e.g. educational examination results, medical conditions or tax records). Administrative data are especially powerful for collecting information that is more difficult to collect with a high degree of accuracy in a social survey context (e.g. the exact start and end dates of a job). In addition the linkage of these made and found data resources will greatly reduce the burden on survey respondents.

2- ADMINISTRATIVE DATA AT IBGE

STRATEGIC MAP FOR CAPACITY BUILDING THROUGH CENSUS



ADMINISTRATIVE Data Registries



Data Quality Control:

- Data sanitization
- Imputation of “missing” data
- Data check

Sampling:

- Master Sample of Household Surveys
- Special Household Samples
- Economic Surveys Samples

Privacy and Confidentiality:

- Anonymization techniques
- Restricted Data Access Room (SAR)
- Evaluation Committee for Access to Non-Identified Microdata (CAD)

1st use

Linkage of different data registries

This type of linkage allows an increase in the dimension of analyzed variables, in addition to the possibility of transforming cross-section information into longitudinal or temporal information.

Matching at the level of companies (via CNPJ) or individuals (via CPF as a Social Security Number), or block (name, date of birth and address).

2nd use — Linkage of Census and household surveys with data registries

Examples:

- Pair the 2013 PNS with the SIM and the civil mortality registry, allowing lifestyles, alcoholism, smoking to be assessed as factors associated with death and specific causes of death.
- Pair the Continuous PNAD with the GFIP, to identify whether a particular family and educational structure is a close determinant for the transition and duration of formal employment.
- Pair PNAD Continuous with GFIP and Social Security, to determine the transition to retirement and pension benefits.

3rd use — Pairing of data registries and Census data

This type of exercise allows the identification of the coverage of public policies in the various spheres, once the Demographic Census generates the denominator (target audience) of the events counted in the records, allowing the calculation of rates.

Longitudinal health survey

Data linkage between National Health Survey data *(Pesquisa Nacional de Saúde – PNS)* 2013 and 2019 and mortality rates

Abordagem da rede de pesquisas

Longitudinal health surveys like those developed by CDC (Centers for Disease Control and Prevention)

3- GEOSPATIAL STATISTICAL FRAMEWORK (GSF)

Open data

Big Data

Statistical Community

Socio-Economic Datasets

Core Statistical
Census, Demographics, Agriculture, Building, Labour Force, etc.

Tax
Income and Business Tax

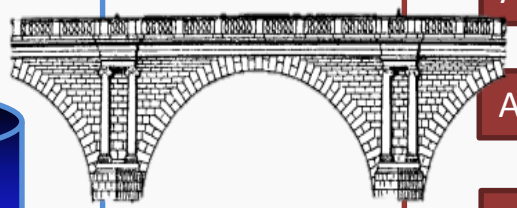
Immigration

Health
Medicare, Pharmaceuticals, Workforce

Land
Valuation and Use

Social Welfare
Unemployment, Disability, Family Support

Others …

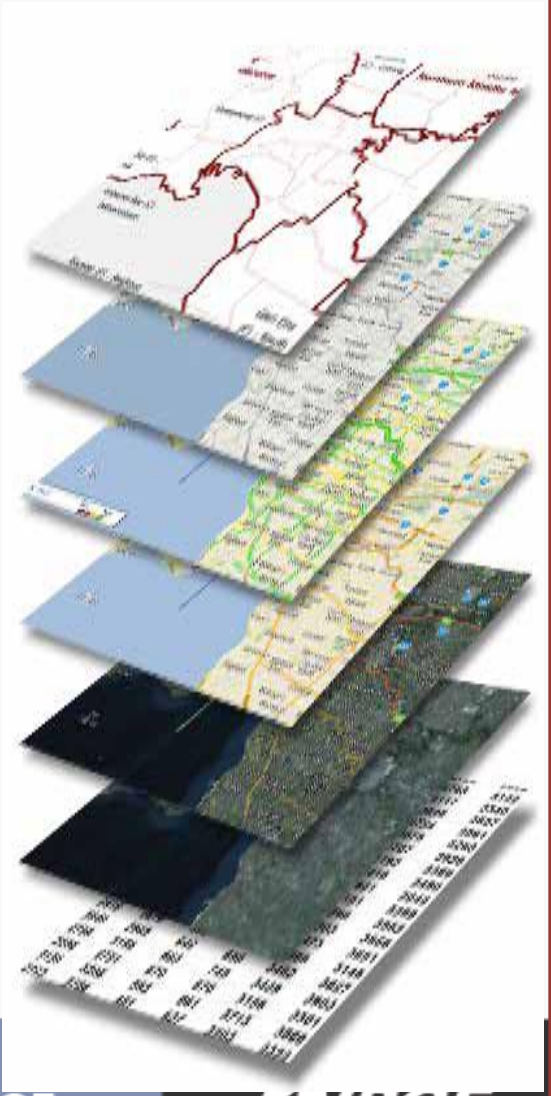


SSF bridge

Spatial Community

Foundation Spatial Data Framework – Fundamental Elements

- Admin. & statistical boundaries
- Addressing, Place Names
- Transport, Water
- Land and Property
- Elevation and Depth
- Imagery
- Positioning



Remote sensing for geospatial and statistical data

Improve the extraction of geospatial information and statistics from remote sensing, with increased automation

Context

Wide availability of **satellite images**

Large-scale image processing capability (cloud processing)

Image processing algorithms (Machine Learning, Artificial Intelligence)

Improvement of the production of geospatial information and statistics from satellite images

4- OTHER BIG DATA ANALYSIS: WEBSCRAPPING

Webscraping and Big Data

Inflation rate (IPCA) fully online, and partly automatized (webscrapping)

Progress made:

- ✓ Further characterization of the problem of extracting web data for for the sample of outlets and products contained in the SNIPC data basis
- ✓ Initial analysis reveal a lower bound of 25% and an upper bound of 80% of data that can be found through the web
- ✓ Based on such analysis, it is being developed a **system** to extract and treat the desired information

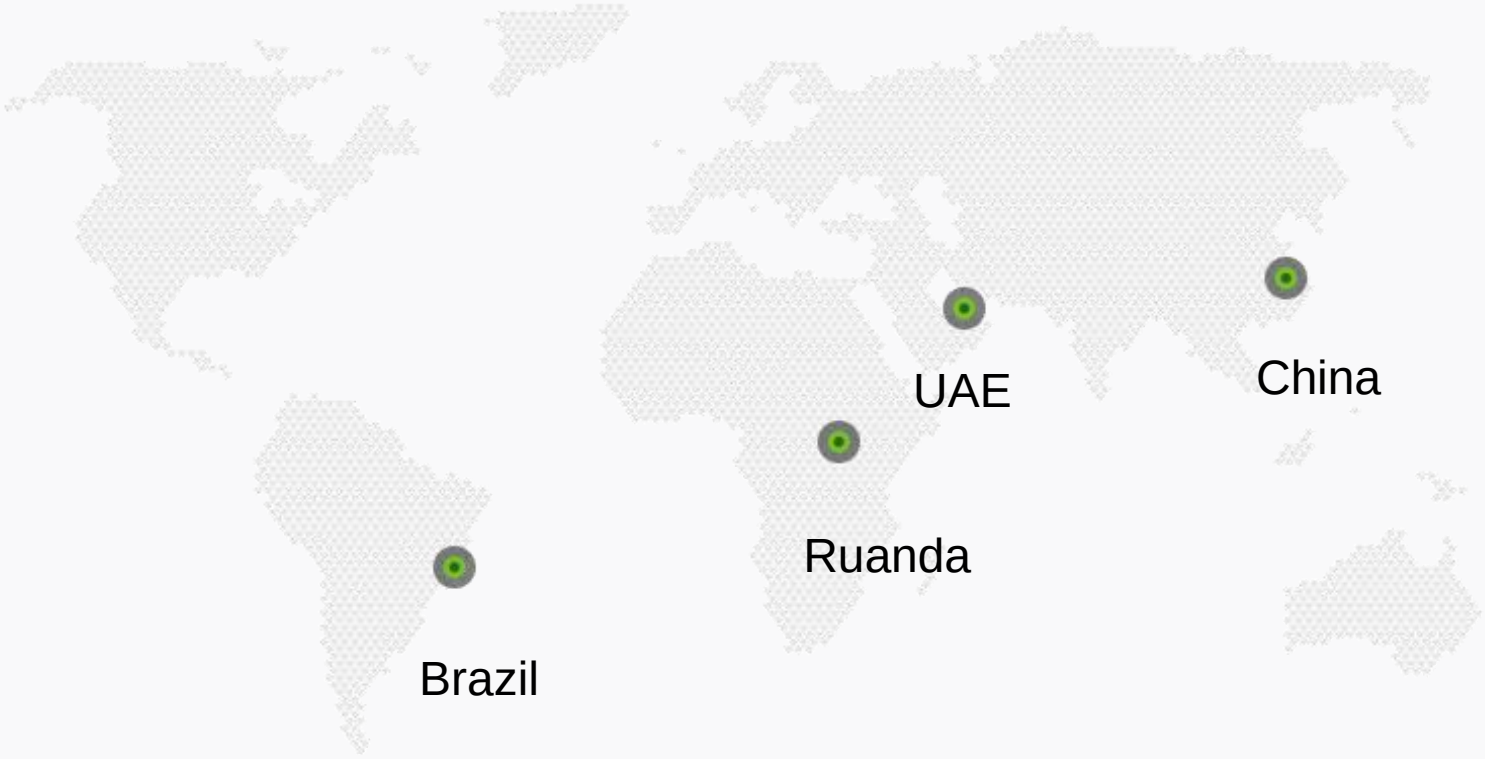
Webscraping and Big Data

The system comprises several modules:

- **Source discovery**
  - designed to find candidate sites
- **Inter-source discovery**
  - designed to find desired information in the sites elected.
- **Intra-source aggregator**
  - aggregates data from the inumerous source in agreement with SNIPC's classification structure
- **Evaluation module**
  - allows evaluation of the system performance
- **Control module**
  - to be used by IBGE staff for data validation and fine tuning of the system

5- REGIONAL HUB FOR BIG DATA IN BRAZIL

Regional Hubs for Big Data expand the platforms reach




UNITED NATIONS | DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS | STATISTICS DIVISION
UNBigData
HOME
ABOUT
EVENTS
TASK TEAMS
UN GLOBAL PLATFORM
Regional Hub for Big Data in Brazil
Rio de Janeiro, Brazil

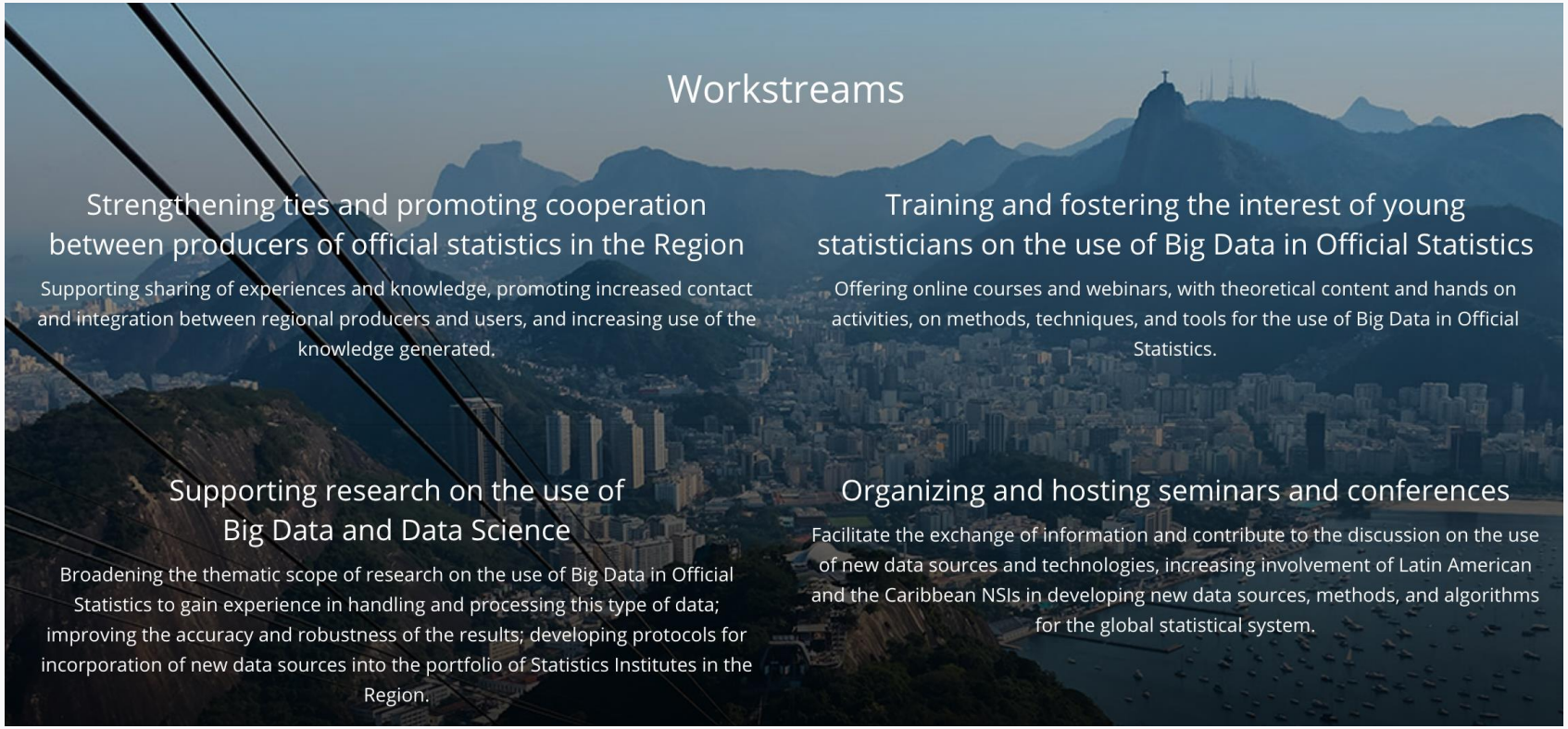
About

The Regional Hub in Brazil aims to contribute to the advancement in the use of Big Data and Data Science to improve the production of official statistics, promoting the sharing of knowledge and the development of innovative initiatives in Latin America and the Caribbean.

About

The Hub is based at the National School of Statistical Sciences (ENCE) of the Brazilian Institute of Geography and Statistics (IBGE).

Founded in 1953, the School has a long tradition of training and research, working at the undergraduate and graduate levels, in addition to offering e-learning.


Workstreams
Strengthening ties and promoting cooperation between producers of official statistics in the Region
Supporting sharing of experiences and knowledge, promoting increased contact and integration between regional producers and users, and increasing use of the knowledge generated.
Training and fostering the interest of young statisticians on the use of Big Data in Official Statistics
Offering online courses and webinars, with theoretical content and hands on activities, on methods, techniques, and tools for the use of Big Data in Official Statistics.
Supporting research on the use of Big Data and Data Science
Broadening the thematic scope of research on the use of Big Data in Official Statistics to gain experience in handling and processing this type of data; improving the accuracy and robustness of the results; developing protocols for incorporation of new data sources into the portfolio of Statistics Institutes in the Region.
Organizing and hosting seminars and conferences
Facilitate the exchange of information and contribute to the discussion on the use of new data sources and technologies, increasing involvement of Latin American and the Caribbean NSIs in developing new data sources, methods, and algorithms for the global statistical system.

Thank you!!!