BigData
UN Global Working Group

4TH INTERNATIONAL CONFERENCE ON

BIG DATA

for Official Statistics

8-10 NOVEMBER 2017

BOGOTA, COLOMBIA

COMBINING DATA SOURCES TO UNDERSTAND THE DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

Marcelo Pitta (Cetic.br) and Denise Silva (IBGE-ENCE)

Bogota | 10th November 2017

GOBIERNO DE COLOMBIA MINTIC DANE INFORMACIÓN ESTRATÉGICA

REGIONAL CENTER FOR STUDIES ON THE DEVELOPMENT OF THE INFORMATION SOCIETY - CETIC.br

Brazilian Internet Steering Committee

cgi.br

Brazilian Network Information Center

nic.br

INTERNET GOVERNANCE

Latin America and Portuguese-Speaking Countries of Africa

1995 — **2003*** — **2005** — **2011** — **2012**

(*) New Federal Decree: creation of NIC.br, changes in the structure and constitution of CGI.br.

cetic.br

United Nations Educational, Scientific and Cultural Organization

cetic.br

Regional Center for Studies on the Development of the Information Society under the auspices of UNESCO

MISSION	METHODOLOGY	POLICYMAKING

- Production and dissemination of ICT statistics
- Promote the use of ICT statistics in policymaking and academic research
- Capacity-building on ICT survey methodologies
- International cooperation for standardization of ICT indicators

UN Economic Commission for Latin America and the Caribbean (CEPAL)

International Telecommunication Union (ITU)

Organization for Economic Co-operation and Development (OECD)

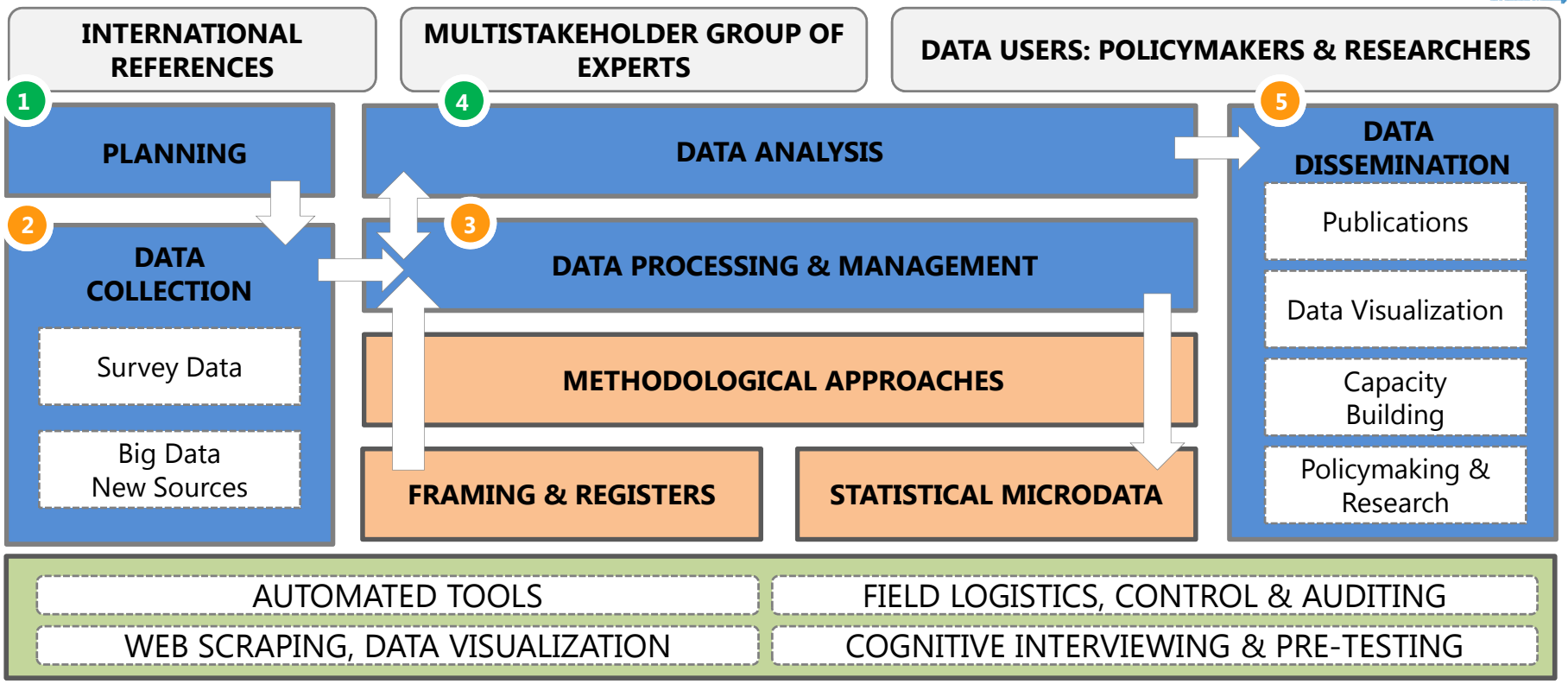
United Nations Educational, Scientific and Cultural Organization (Unesco)

SURVEYS

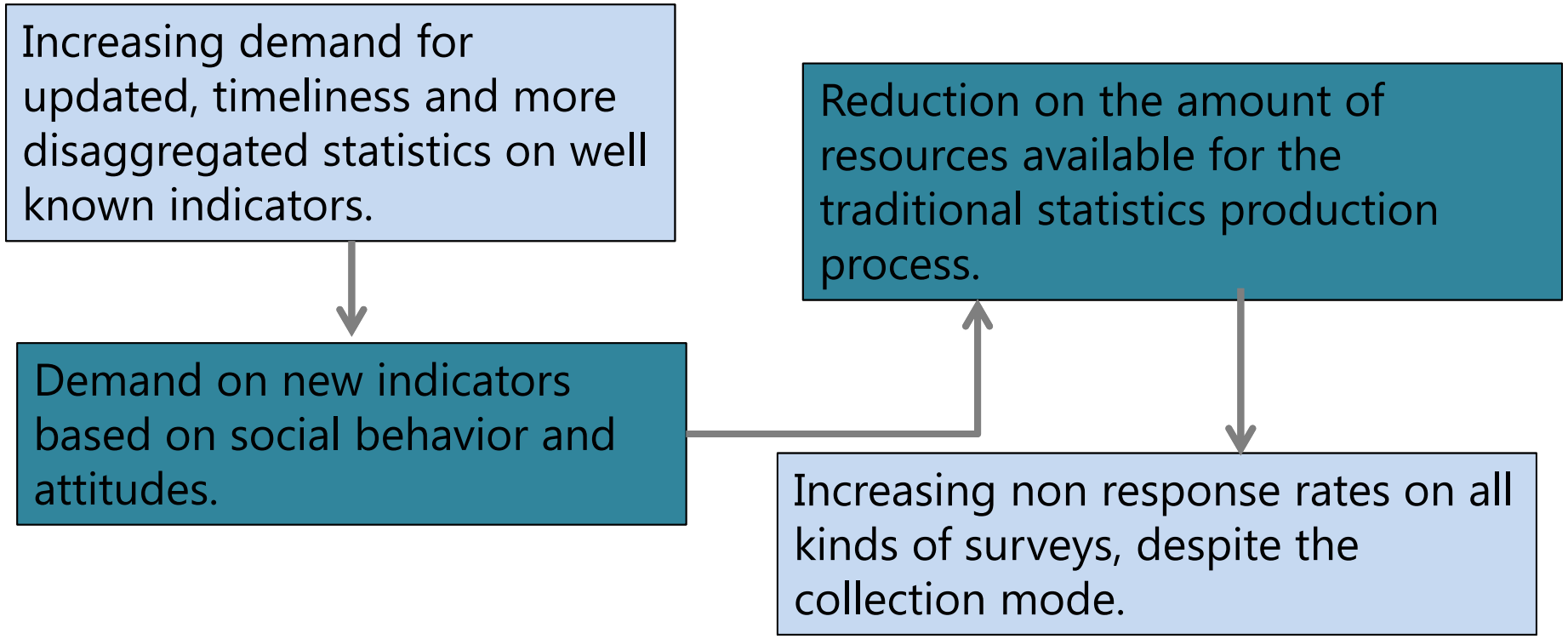
INDIVIDUALS	ORGANIZATIONS		
❑ HOUSEHOLDS & INDIVIDUALS	❑ EDUCATION	❑ ENTERPRISES	❑ ISP PROVIDERS
❑ RIGHTS & PROTECTION	❑ HEALTH	❑ GOVERNMENT	❑ TELECENTERS
	❑ CULTURE	❑ NON-GOV'T ORG.	

ICT STATISTICS PRODUCTION PROCESS

PROCESS AND PLATFORM



THE SCENARIO FOR THE PRODUCTION OF STATISTICS IS CHANGING…



COMBINING DATA SOURCES

SURVEYS AND ORGANIC (BIG) DATA

A NEW SOCIOECONOMIC ECOSYSTEM SELF-MONITORED IS EMERGING

DESIGNED (TRADITIONAL SURVEY) **DATA**

- *DATA PRODUCED IN STRUCTURED FORM*

+

ORGANIC (OR BIG) **DATA**

- *DATA PRODUCED FROM AUXILIARY PROCESS AND EVENT MONITORING*

COMBINING THESE TYPES OF DATA IS THE FUTURE

DR. ROBERT GROVES: PROVOST OF GEORGETOWN UNIVERSITY – WASHINGTON D.C. AND FORMER DIRECTOR OF US CENSUS BUREAU

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

PHASE 1

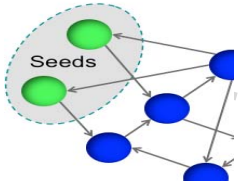
BUILDING A PREDICTION MODEL

Web scraping process
Survey Data
Semantic Dictionary

- website1.com.br
- website3.com.br
- website4.com.br
- . . .
- website*n*.com.br

SEMANTIC DICTIONARY

Prediction Model



PHASE 2

PRODUCING ICT INDICATORS

Sampling (Domain Name System Frame) (DNS)
Web data collection
Statistical Model

DNS DATABASE

Domain Name Database

Prediction Model

Web Scraping Process

Data Analysis & Report

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

- **OBJECTIVE:**
  - Produce selected ICT indicators using automated data collection tools (web scraping on web pages)
- **MOTIVATION:**
  - Increase the overall sample size of traditional surveys
  - Reduce response burden
- **EXPECTED OUTCOMES:**
  - Accuracy evaluation of statistical models based on big data sources to estimate ICT enterprise indicators
  - Development of tools for automated data collection, classification and data modeling in the Web

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

- **MODELING:**
  - *Fit logistic and multinomial logistic models based on 2015/2016 survey data to predict selected indicators*
- **SELECTED INDICATORS:**
  - *Proportion of enterprises that offer in their websites* - Product catalog, Price list, Ordering system, Online payment, Client support
  - *Proportion of enterprises that purchase on the Internet*
  - *Proportion of enterprises that sell on the Internet*
  - *Proportion of enterprises that sell on the Internet through e-mail*
  - *Proportion of enterprises that sell on the Internet through social networks*
  - *Proportion of enterprises that sell on the Internet through group buying sites*

THE ICT ENTERPRISES SURVEY

METHODOLOGICAL ISSUES

- **SAMPLE SIZE:**

  7.000 enterprises with 10 or more employees:

  - All geographical regions of Brazil;
  - Size: small, medium and large enterprises;
  - Economic activity: 11 sectors ISIC 4.0

- **INFORMATION UNIT:**

  IT or network professionals

  - Large companies: second respondent from the Accounting and Legal Departments

- **METHOD OF DATA COLLECTION**

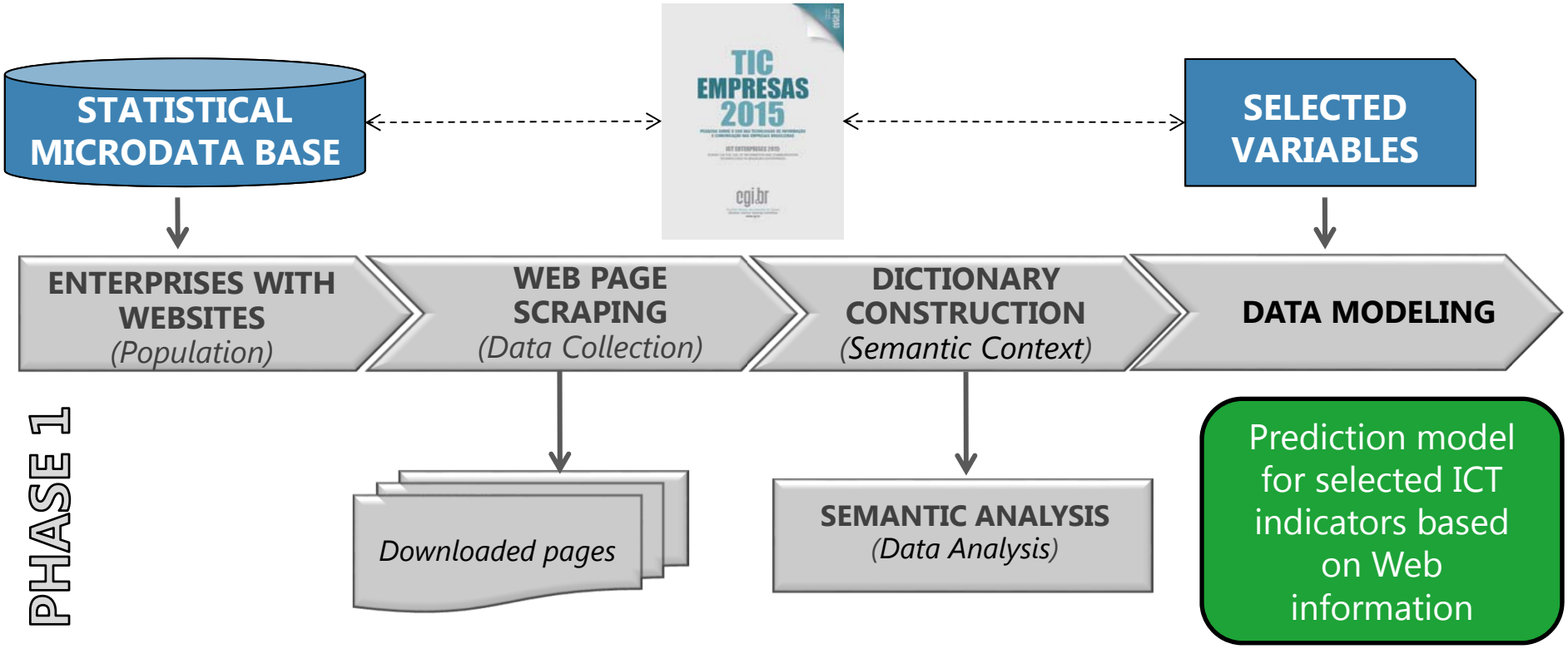
  Interviews by phone (CATI)

- **DATA COLLECTION PERIOD:**

  From Sept/2015 to Dec/2015

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES



BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

Logistic model for complex survey data that takes into account the survey design

$$Y = \begin{cases} 1, \text{if the enterprise operates Internet selling} \\ 0, \text{otherwise} \end{cases}$$

$$X \Rightarrow \text{data collected via enterprise } \textit{webpages}$$

$$Log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta X$$

PHASE 1

ICT Enterprise survey data

web scraping data

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

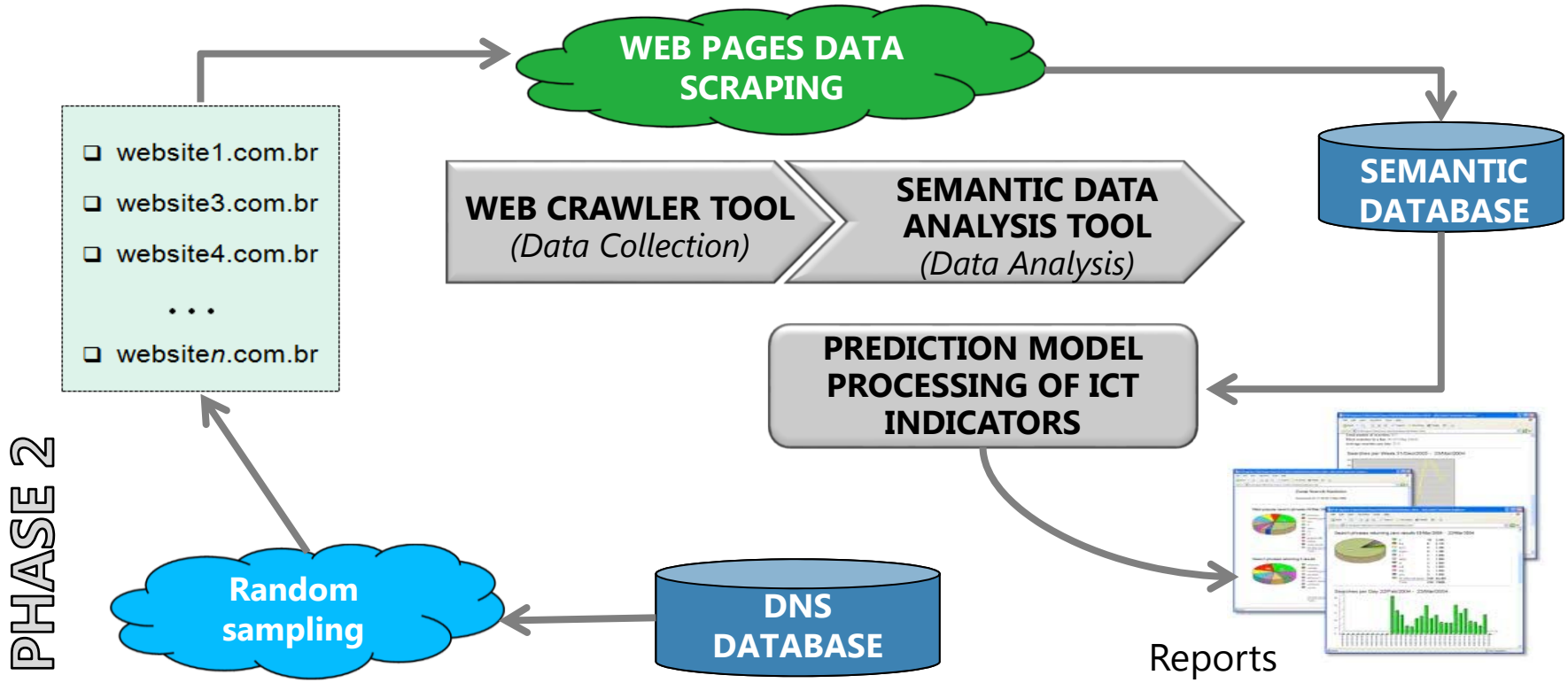
Phase 1 Challenges

- Use of alternative modeling procedures that can take into account complex survey designs; and

- To build a tool to automatically address and create the *X* variables for the model (still requires intense human intervention).

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES



BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

Phase 2 Challenges

- Access to the DNS and Enterprises database in a regular basis;
- Capture the changes in the words and the way enterprise *web pages* function; and
- Automatically identify the changes and promote modeling adjustments (development of a tool).

BIG DATA PILOT PROJECT ON DIGITAL ECONOMY

USING WEB SCRAPING TO PRODUCE ICT INDICATORS FOR ENTERPRISES

Solutions to address the challenges:

- New methods are being developed to deal with complex sample surveys;
- Semantic analysis is an option to deal with the automation of data collection for modeling purposes in phase one;
- Panel survey of enterprises, comparing and following the changes over time in the web pages to evaluate differences in the semantic and the structure of ~~the~~ pages; promoting model revision;
- Updating the model every two years based on the Enterprise ICT survey data.

THE UNITED NATIONS GLOBAL
WORKING GROUP ON BIG DATA

Thank you for your attention!

Muchas gracias por su atención!

www.cetic.br

#UNBIGDATA2017

GOBIERNO DE COLOMBIA MINTIC DANE INFORMACIÓN ESTRATÉGICA