United Nations Economic Commission for Europe
Statistical Division

modernstats

Workshop on the Modernisation of Official Statistics
November 24-25, 2015

The Sandbox project

The Sandbox 2015 Report

Antonino Virgillito
Carlo Vaccari

Project coordinators, UNECE - ISTAT

INTRODUCTION

ACTIVITIES

FUTURE OF SANDBOX

OUTCOMES

Our presentation

- Introduction
- Activities
  - Comtrade
  - Enterprise Web Sites
  - Wikistats
  - Twitter
- Future
- Conclusions

What is the Sandbox

Shared computing environment developed in partnership with the Irish Central Statistics Office and the Irish Centre for High-End Computing (ICHEC)

A unique platform where participating organisations can engage in collaborative research activities

Open to all producer of official statistics

Sandbox background: 2014

Big Data project 2014: four Work Packages

1 Work Package 1: Issues and Methodology

2 Work Package 2: Shared computing environment ('sandbox') and practical application

3 Work Package 3: Training and dissemination

4 Work Package 4: Project management and coordination

- Social Media
- Mobile Phones
- Prices
- Smart Meters
- Job Vacancies
- Web scraping
- Traffic Loops

Sandbox 2015 goals

Publish a set of international statistics based on Big Data, before the end of the year

Choose 2-3 Big Data sources

Collect Data

Compute multi-national statistics

Organize press conference

Conclude 2014 experiments on the Sandbox

The Sandbox 2015

Installed Software

— Hadoop (Hortonworks Data Platform)
— R – Rstudio new
— RHadoop
— Spark new
— ElasticSearch new

and growing…

4 Data/Compute nodes new
— 2 x 10 core Intel Xeon CPUs
— 128 GB RAM
— 4 x 4TB disks
— 56 Gbit InfiniBand network

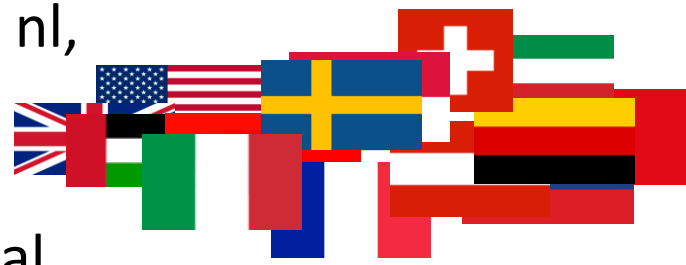
2 Service/login nodes new
— Similar hw as data nodes
— 10Gbit connection to Internet

Participants

More than 40 people from 22 entities

International organisations: Eurostat, OECD, UNECE, UNSD

Countries: at, ch, de, es, fr, hu, ie, it, mx, nl, pl, rs, ru, se, si, tr, uae, uk, us



Active participants: about half of the total

ICHEC
Irish Centre for High-End Computing

ICHEC Assisted the task team for the testing and evaluation of Hadoop work-flows and associated data analysis application software

Work Groups

Four main activities, for each task one multinational group

Wikistats - Wikipedia hourly page views: use of an alternative data source

Twitter - Social media data: experiences comparison in tweets collection and analysis

Enterprise websites: the Web as data source - web scraping and business registers

Comtrade - UN global trade data: use of Big Data tools on a traditional data source

INTRODUCTION

ACTIVITIES **Wikistats**

FUTURE OF SANDBOX

OUTCOMES

Description

Statistics on hourly page views of articles in the several language versions of Wikipedia

Download and process to obtain manageable data

1. UNESCO heritage sites
2. Cities and Beaches

Popularity
Touristic potential

Wikipedia page views is a potential source for many domains: tourism, culture, economic, …

Data Characteristics

Wikipedia
seventh
website (Alexa)

Public source
(contents and metadata)

Digital traces
left by people in
their activities

Widely used:
44% of EU 16-74
69% of EU 16-24

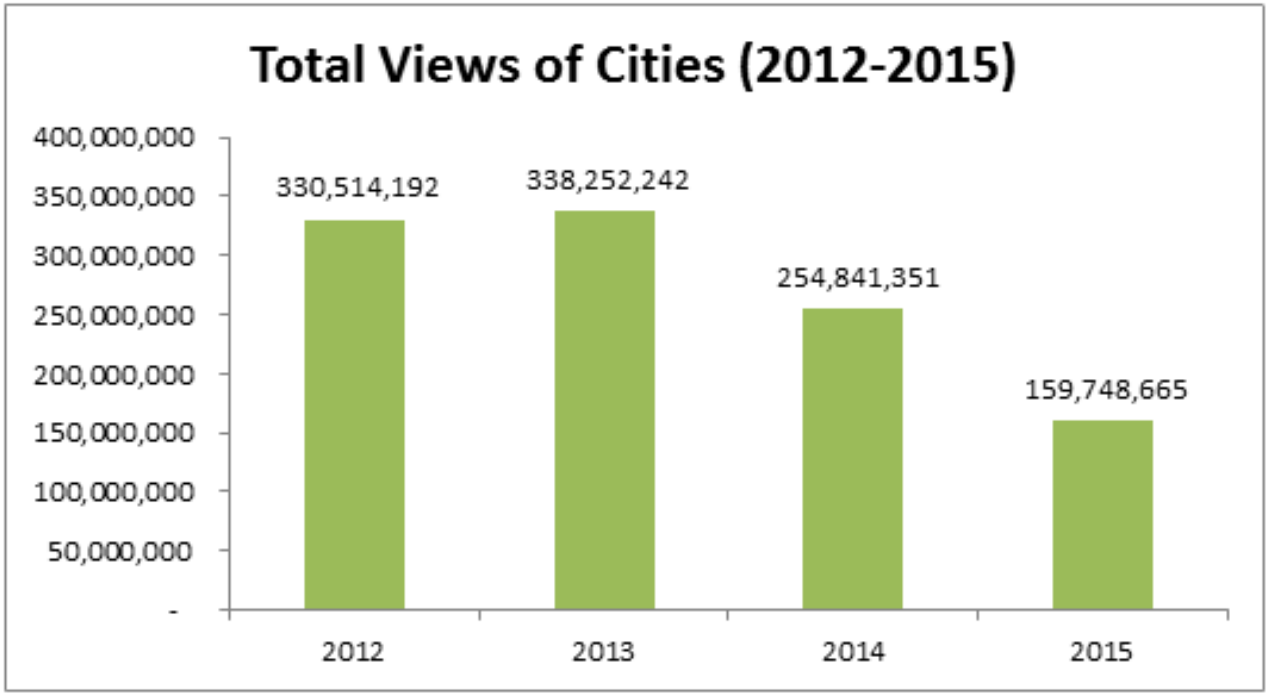
Activities

Pre-processing: scripts in shell and Pig to filter data and change time-series format

Extraction: MapReduce, shell and python to filter articles and time aggregation (hourly to daily, weekly and monthly)

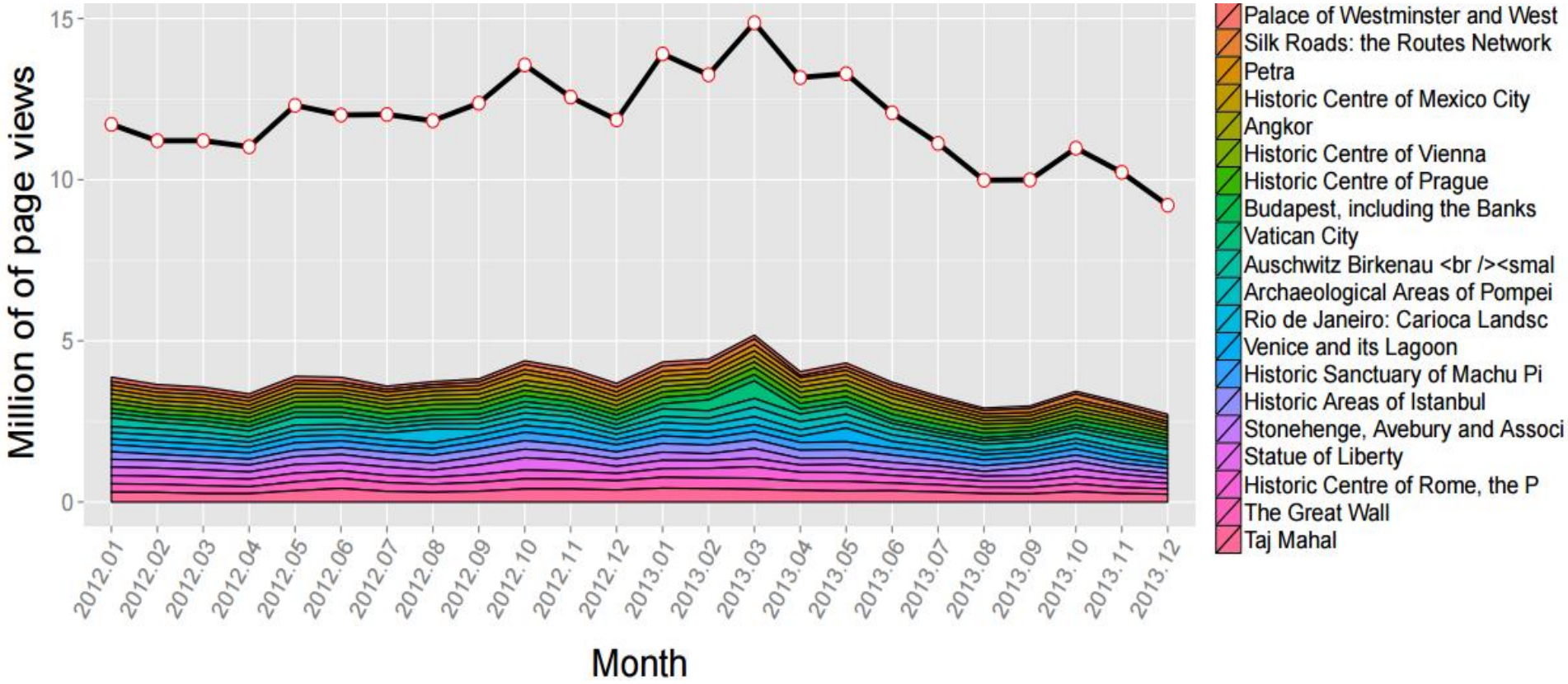
Analytics: R and RStudio for web scraping, selection of articles and Data analysis

Outputs



More than one billion views!

Outputs



See peak for Vatican City in 2013 March for pope Francis election

Findings

Relevance:
Good for culture and regional statistics
New topics uncovered so far
Many other potential topics
Unprecedented temporal detail

Technology:
Experience: scripts, Pig and RStudio
Try NoSQL, Spark, Elasticsearch
Problems with Java integration
Need resource-scheduling system

Findings

Source

Good potential, need more investigation
Data available, adequate IT needed
No privacy concerns also on individuals
Continuity: substantial, not guaranteed
More mobile users: need to investigate

Quality

Accuracy (bot excluded), categorization
Timeliness: good, data few hours later
Comparability in space and between languages
Crowd-sourced: improve completeness
Can improve statistics: new phenomena

How far are we from something that can be published?

a) Concerning the results of the experiment

- Almost ready for publish first experimental statistics with 30 languages (Issues on encoding)

b) Concerning statistics based on this data source in general

- Results data source well received by domain experts (page views)
- No issues of accuracy, but interpretability issues need validation
- New statistics like ”index of consumption of digital cultural products” require methodological development and conceptual frameworks

What case has been made for the future of the Sandbox?

- Sandbox fundamental to experiment with this data source (desktop computers are simply not able to process this source)
- Present and future existence of the sandbox is important to develop further the work done, with tools which can increase the process efficiency (HBase Elasticsearch)
- Sandbox is also important to make the system available for other statisticians: the code is available in GitHub, but not the data. In the sandbox, where the data is available, statisticians can re-use the code, try and build other applications on those data.

INTRODUCTION

ACTIVITIES **Twitter**

FUTURE OF SANDBOX

OUTCOMES

Description

Geolocated tweets

2015: two approaches:

- Public stream (Mexico , UK)
- Buy data (UK)

Source instability with technological changes

Comparison of Data Preprocessing

Data analysis: distinct targets, methods and tools

Data characteristics

	UK ONS	Mexico INEGI
Period	April - August 2014 (API) August - October (gnip)	February 2014 – Today (running)
Method	Public Twitter API Purchased from GNIP	Public Twitter API
Percentage of Twitter Active Users	5.6%	3.0%
Percentage of Geo-Referenced Tweets	1.57% (London)	1.03% (Mexico City)
No-SQL Database	CSV, MongoDB	Elastic Search
Total Records:	~106 million collected ~81.4 million used	~150 million collected ~74 million used

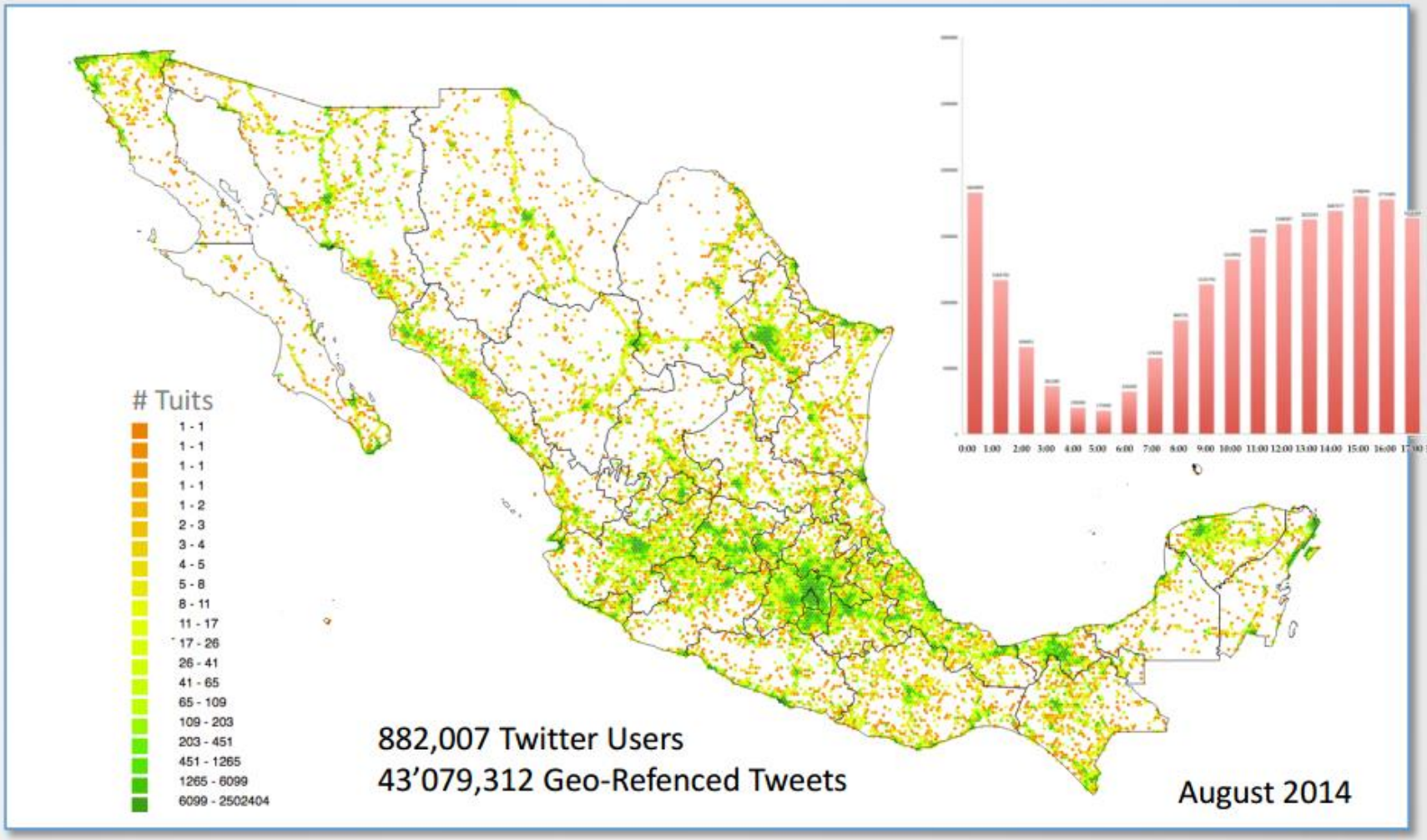
Outputs

Procedures for collecting geolocated tweets from public stream:

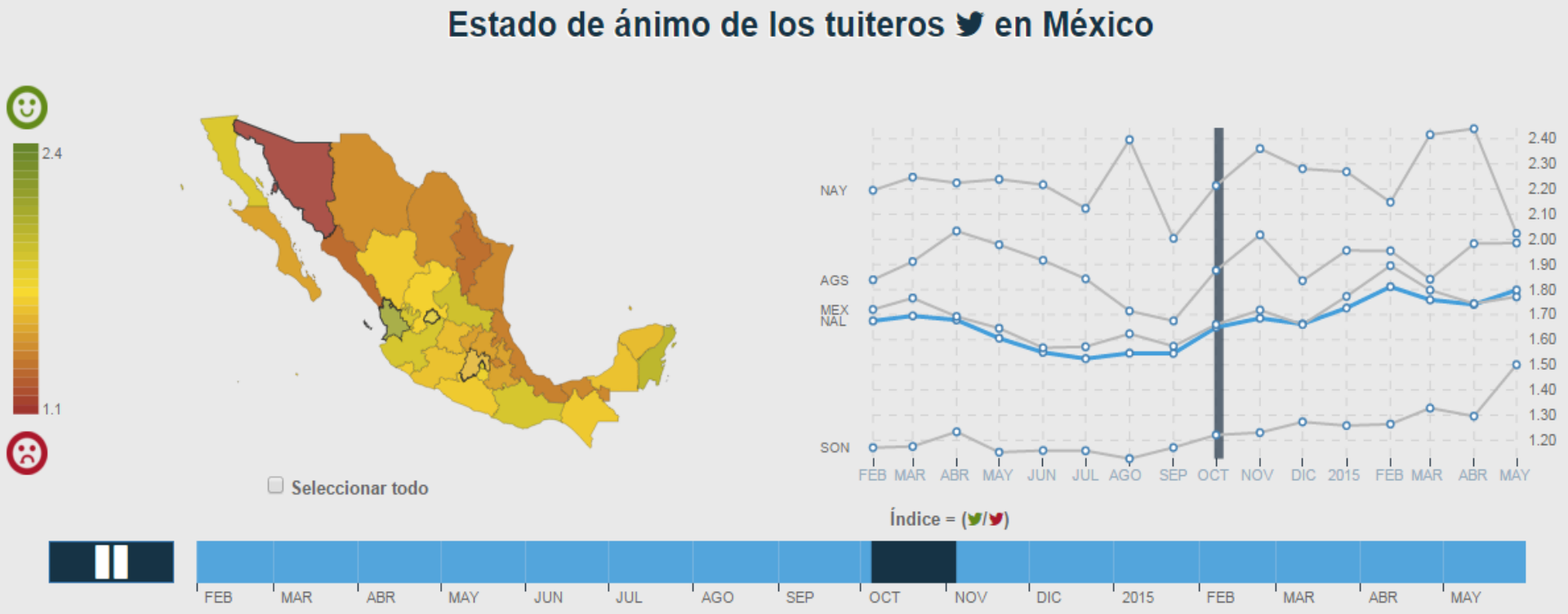
- Mexican version
- Sandbox version

Data Collection in progress on the Sandbox for tweets geolocated in Rome (Italy) for future study on Jubilee 2015-2016

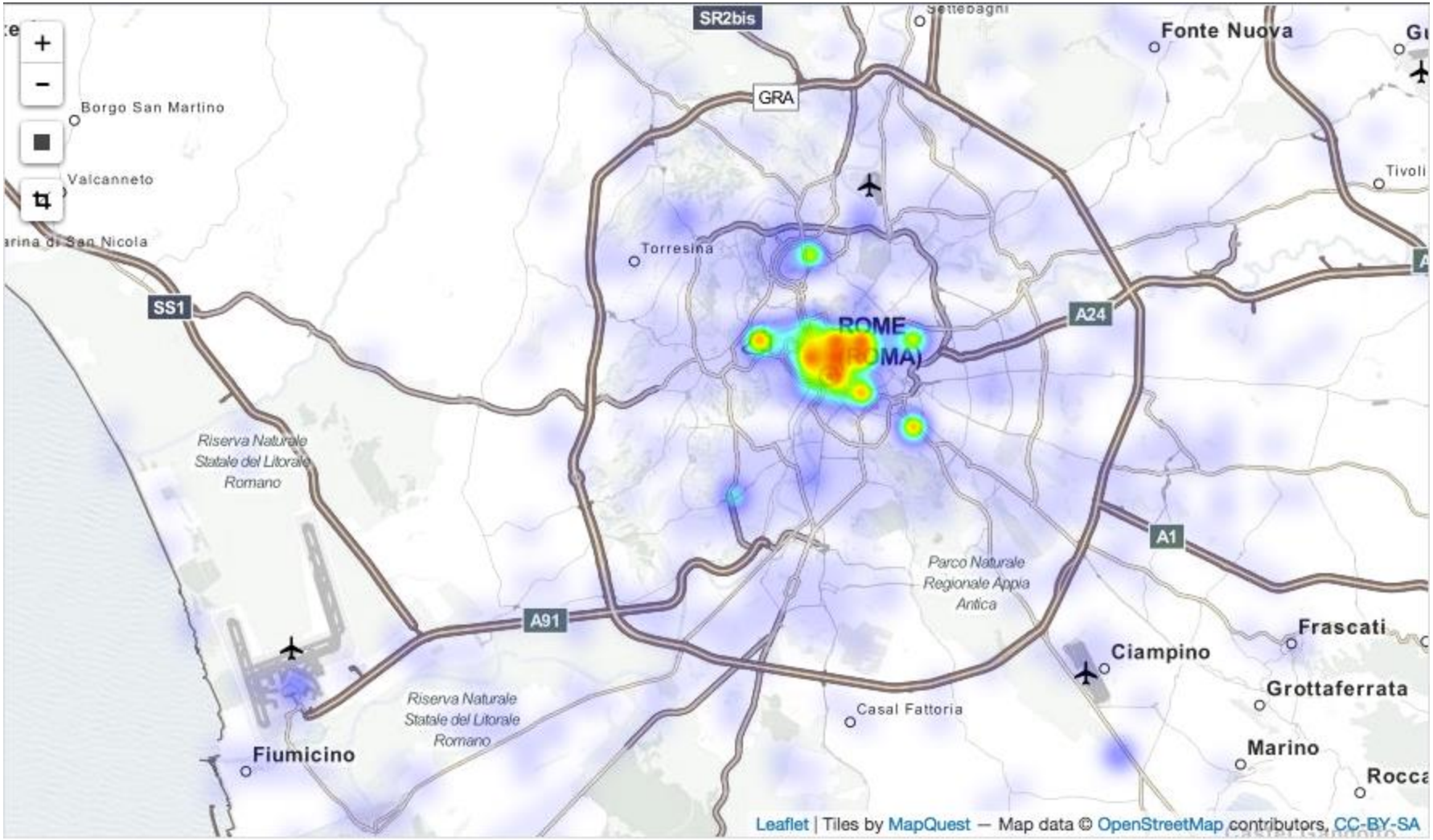
Outputs


Tuits
1 - 1
1 - 1
1 - 1
1 - 1
1 - 2
2 - 3
3 - 4
4 - 5
5 - 8
8 - 11
11 - 17
17 - 26
26 - 41
41 - 65
65 - 109
109 - 203
203 - 451
451 - 1265
1265 - 6099
6099 - 2502404
882,007 Twitter Users
43'079,312 Geo-Refenced Tweets
August 2014

Outputs

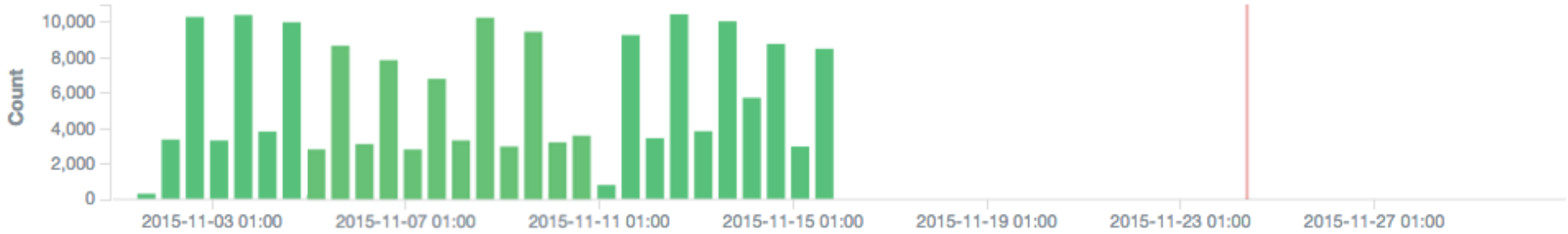

Estado de ánimo de los tuiteros en México
2.4
1.1
Seleccionar todo
NAY
AGS
MEX NAL
SON
2.40
2.30
2.20
2.10
2.00
1.90
1.80
1.70
1.60
1.50
1.40
1.30
1.20
FEB MAR ABR MAY JUN JUL AGO SEP OCT NOV DIC 2015 FEB MAR ABR MAY
Índice = (/)
FEB MAR ABR MAY JUN JUL AGO SEP OCT NOV DIC 2015 FEB MAR ABR MAY

Outputs

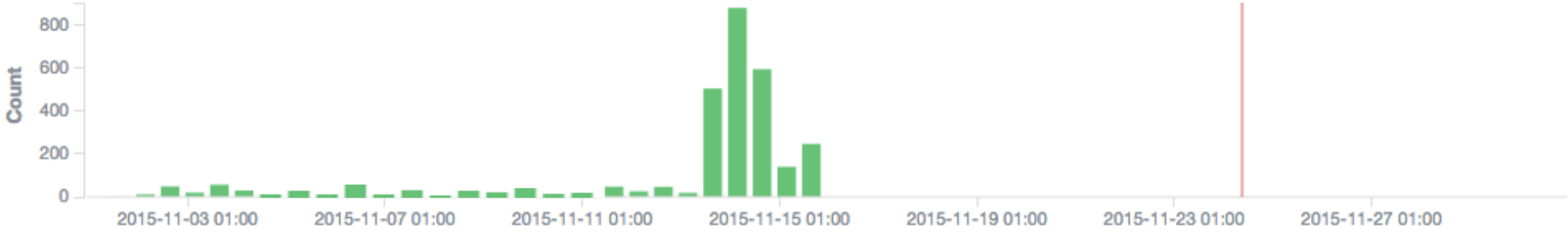

SR2bis
GRA
Fonte Nuova
Borgo San Martino
Valcanneto
Tivoli
Torresina
SS1
A24
ROME
(ROMA)
Riserva Naturale
Statale del Litorale
Romano
A1
Parco Naturale
Regionale Appia
Antica
A91
Frascati
Ciampino
Grottaferrata
Casal Fattoria
Riserva Naturale
Statale del Litorale
Romano
Fiumicino
Marino
Leaflet | Tiles by MapQuest — Map data © OpenStreetMap contributors, CC-BY-SA

Outputs

Number of tweets generated in the Rome area (each bar = 12 hours)



Filtered on tweets containing the words “Paris” or “Parigi”



Findings

Technology

Elastic (Elasticsearch) good tool to collect and index data

Kibana is a good tool to visualize data

Source

Continuity: technology outside NSOs control

Continuity depends on even small changes to technology

Rules for Data collection: fragile legal basis, better acquire data (cost UK £25,000/year)

How far are we from something that can be published?

- For people mobility we are close to “official usage” in Mexico
- Sentiment analysis Twitter data quite officially used in Netherlands

What case has been made for the future of the Sandbox?

- Procedures to collect data from Twitter ready on the Sandbox, together with tools ready to analyze those data
- Data collection already active on Rome geolocated tweets, ready for analysis on tourists mobility during Jubilee

INTRODUCTION

ACTIVITIES

Enterprise Web Sites

FUTURE OF SANDBOX

OUTCOMES

Description

Objective: using web sites of enterprises as a source to create statistics

Issues

Obtaining URLs of web sites

Different approaches were experimented

Obtain from registries/survey

Obtain from search engines

Implementing a method for scraping data from web sites

Different approaches were experimented for both scraping and analysis phases

Computing the statistics

Number of enterprises which advertise Job Vacancies
broken down by NACE activity and region

Data Characteristics

	Sweden	Slovenia
Websites	13.000	15.000
Websites used in the experiment	1.500	1.335
Web pages	50.000	**1.068.000**

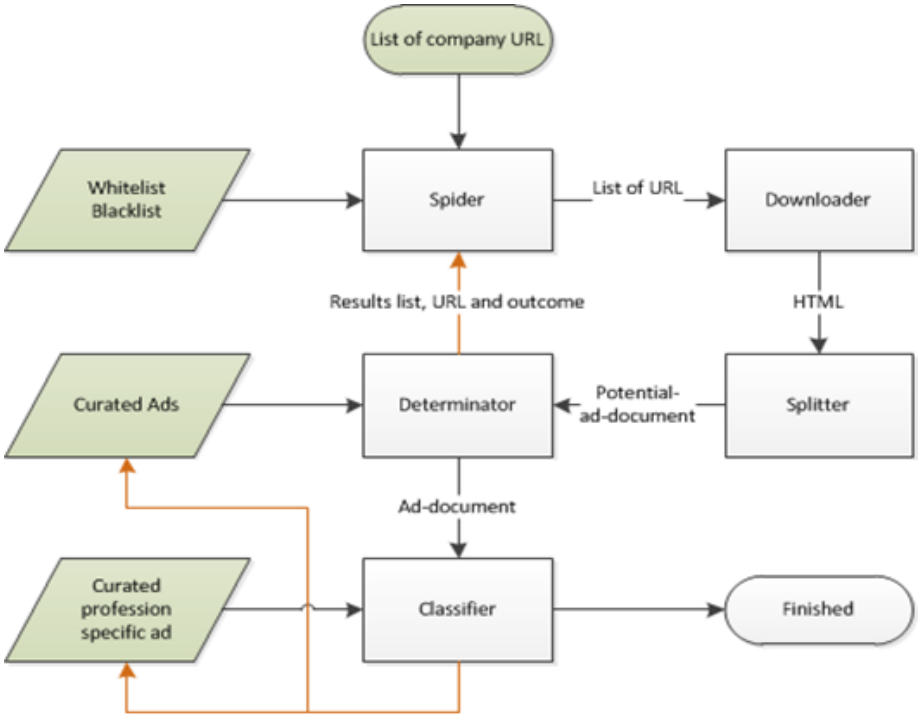
Activities

Development of an application for scraping the job vacancies data starting from URLs of enterprises

- Not tied to language/method
- Developed in Python
- Available in the Sandbox

Spider – Downloader
Start from a list of URLs, follow the links and download the content of the employment pages

Determinator – Classifier
Implement a method for detecting and classifying job vacancies advertisements in the scrapped content



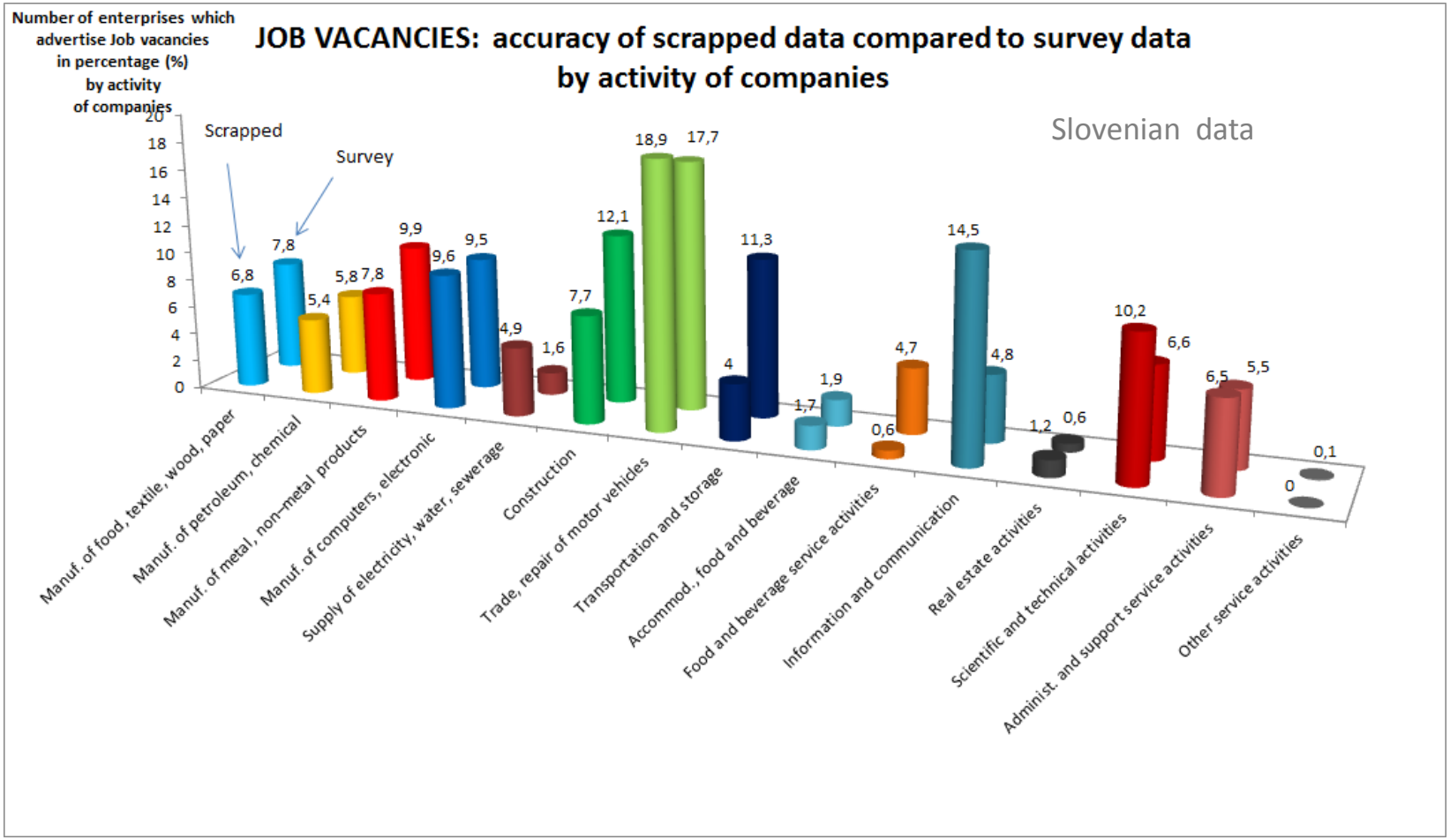
Activities

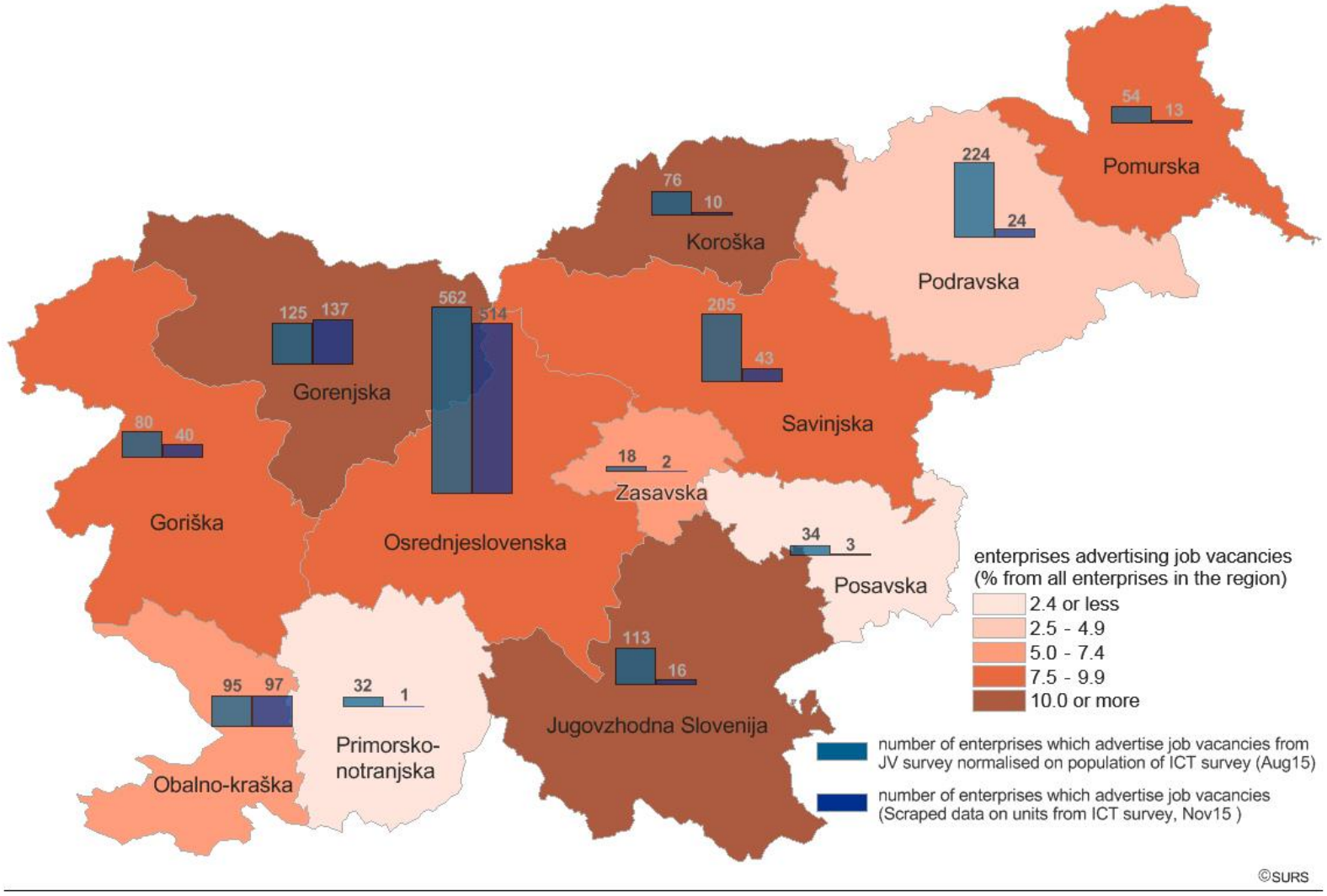
- Developed two different methods for implementing the Determinator module (detection of job vacancies)
  - machine learning approach: 92% accuracy (Swedish data)
  - Keyword (phrases)-based: 80 % accuracy (Slovenian data)
- Created statistics from scraped data
- Tested a technological stack for scraping/analyzing web sites on a large scale
  - Nutch – Elasticsearch - Kibana

Outputs

- Program to scrape and analyze web sites
  - deployed in the Sandbox, can be used in general
- Statistics about the number of JV per NACE group, as calculated from the scraped data
  - Distribution compared with that obtained from the survey
  - Distribution using survey weights compared with distribution using calibrated weights (by number of employees)

Outputs


JOB VACANCIES: accuracy of scrapped data compared to survey data by activity of companies
Number of enterprises which advertise Job vacancies in percentage (%) by activity of companies
Slovenian data
Scrapped
Survey
0
2
4
6
8
10
12
14
16
18
20
6,8
7,8
5,4
5,8
7,8
9,9
9,6
9,5
4,9
1,6
7,7
12,1
18,9
17,7
4
11,3
1,7
1,9
0,6
4,7
14,5
4,8
1,2
0,6
10,2
6,6
6,5
5,5
0
0,1
Manuf. of food, textile, wood, paper
Manuf. of petroleum, chemical
Manuf. of metal, non-metal products
Manuf. of computers, electronic
Supply of electricity, water, sewerage
Construction
Trade, repair of motor vehicles
Transportation and storage
Accommod., food and beverage
Food and beverage service activities
Information and communication
Real estate activities
Scientific and technical activities
Administ. and support service activities
Other service activities


Pomurska
54
13
Podravska
224
24
Koroška
76
10
Gorenjska
125
137
Osrednjeslovenska
562
514
Savinjska
205
43
Zasavska
18
2
Posavska
34
3
Goriška
80
40
Jugovzhodna Slovenija
113
16
Primorsko-
notranjska
32
1
Obalno-kraška
95
97
enterprises advertising job vacancies
(% from all enterprises in the region)
2.4 or less
2.5 - 4.9
5.0 - 7.4
7.5 - 9.9
10.0 or more
number of enterprises which advertise job vacancies from JV survey normalised on population of ICT survey (Aug15)
number of enterprises which advertise job vacancies (Scraped data on units from ICT survey, Nov15)
©SURS

Findings

Privacy
Retrieving and sharing URLs of web sites was not as easy as expected

Still not clear yet whether URLs collected as microdata from Slovenian survey could be used in the Sandbox

Methodology
Promising results from the machine learning approach to identification of job vacancies

Methodology
Comparison with distribution of job vacancies per NACE revealed coherence with survey results, indicating that the approach is solid

How far are we from something that can be published?

- The main problem for publishing the results are non existent lists of URLs of enterprises in most of the countries (i.e., impossible to determine population)
- As an experimental activity, we started to get first results only recently. Six further months of activity would be needed to reach solid results
- However, the prototype is producing promising outcomes in two countries
- Reliable statistics will most probably be based on multiple sources (Job portals and administrative data from agency of Employment,..)

What case has been made for the future of the Sandbox?

- The prototype of the IT tool for detecting the JV is available in the Sandbox
- It can be used by other countries with small modifications of parameters

INTRODUCTION

ACTIVITIES

Comtrade

FUTURE OF SANDBOX

OUTCOMES

Description

UN Comtrade compiles official trade statistics database since 1962 containing billions of records

Due to interest in measuring economic globalization through trade, trade data has been used to analyse interlink between economies

This project is intended to exploit the capability of the tools in Sandbox to process large amount of data to perform analysis of the trade network on a wider scale

identify regional global value chain networks and analyse their properties.

Data Characteristics

- Dump of the comtrade DB
- Each record represents flow of import export between two countries (reporter and partner)
- The data that is stored in the sandbox are from year 2000 to 2013 in Harmonized System classification for all available reporters.
- The total number of data points is around 325 millions records.

Activities

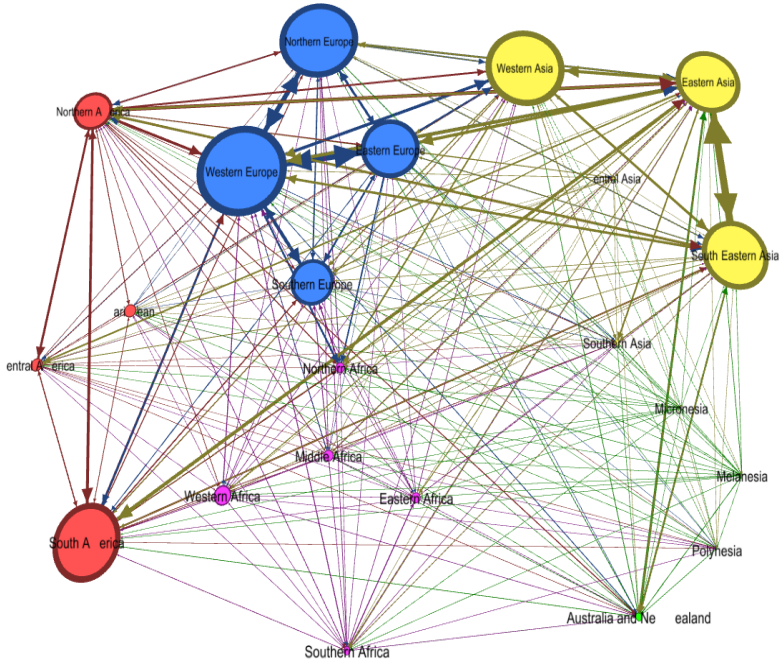
		Tools:
Data acquisition	Extraction from UN Comtrade database, transfer via FTP to the Sandbox. Data loaded into HDFS and made available in Hive Data cleaned up before it could be fed into Hive (eliminate quotes, commas etc.)	Pig Hive
Data preparation	Imputation of gaps, present in the data as not all countries regularly report to UN Comtrade for the combination of reporter-period. Remapping of codes.	Pig Hive

Comtrade - Activities

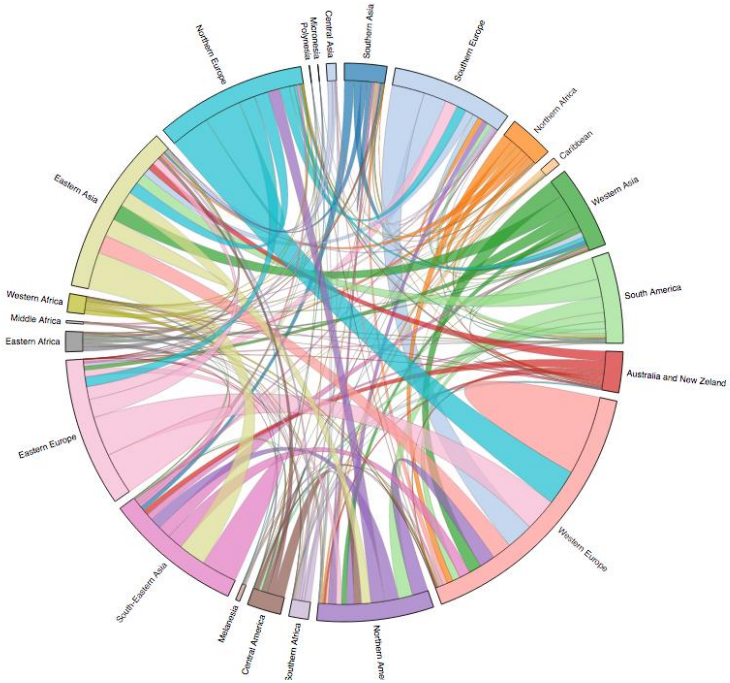
		Tools:
Quality analysis	Measured the coherence between symmetric flows and detect number of missing values. Tested two different approaches	Hive + Python RHadoop
Data visualization	Built several visualizations of the trade networks using different tools	Gephi D3
Network analysis	Computed several metrics of graph characteristics analysis of differences and trends in graphs indicators. Tested two different approaches	Spark Hive + R

Outputs - Visualizations

Trade of intermediate goods by geographical region



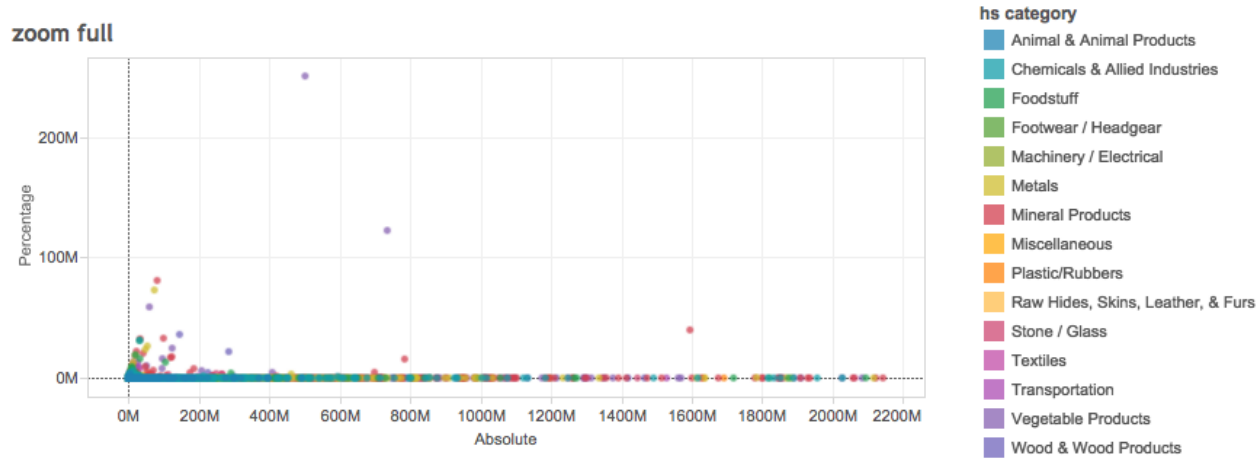
Made with Gephi

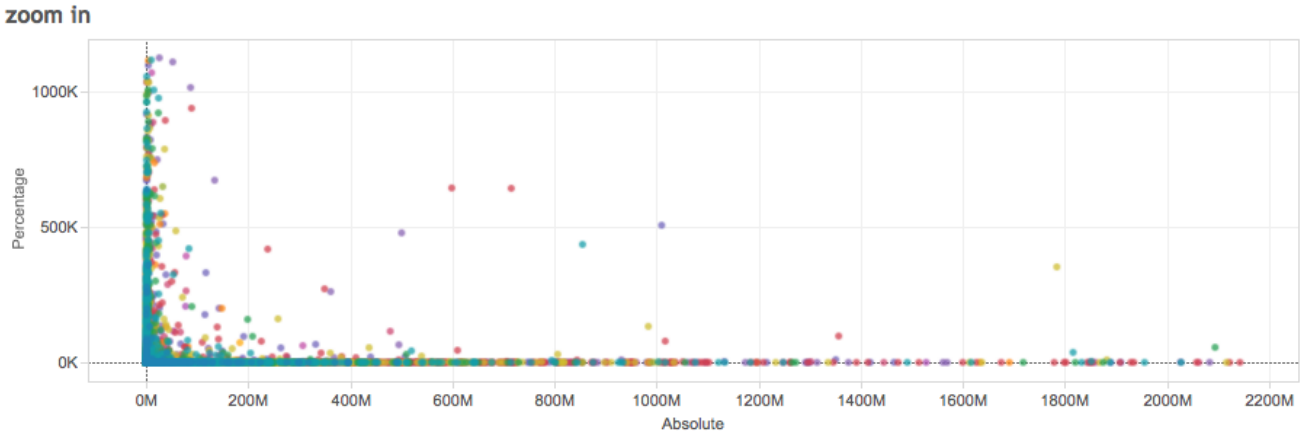


Made with D3

Outputs – Quality Analysis

Differences in symmetric import-export flows (absolute vs percentage)





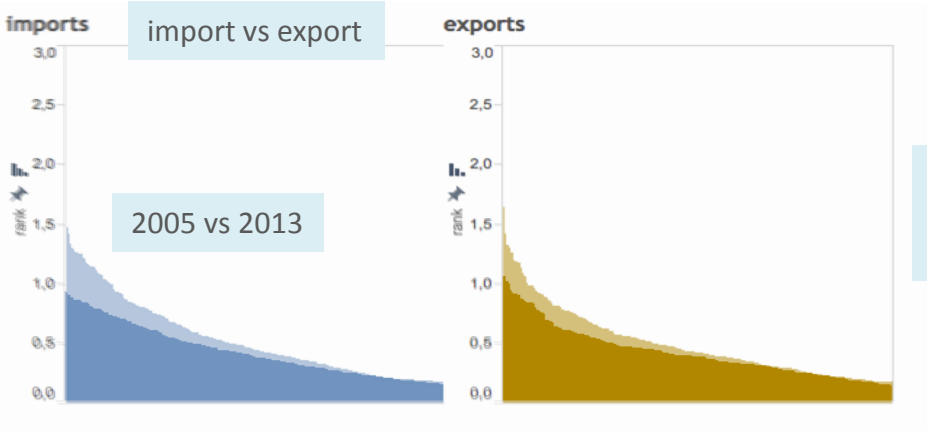
Outputs – Network Analysis

Distribution of PageRank

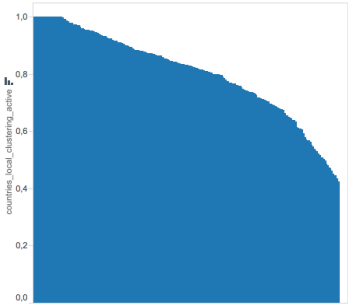
Computed with Spark

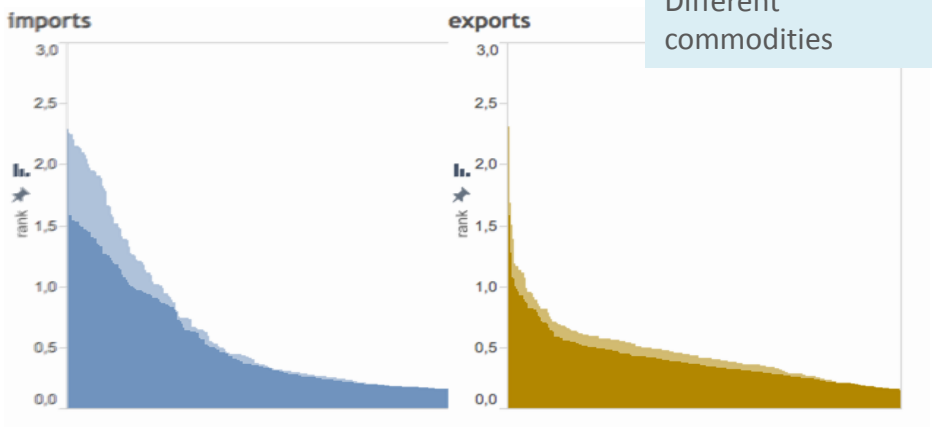
Distribution of local clustering

Computed with Hive + R



Different network indicators

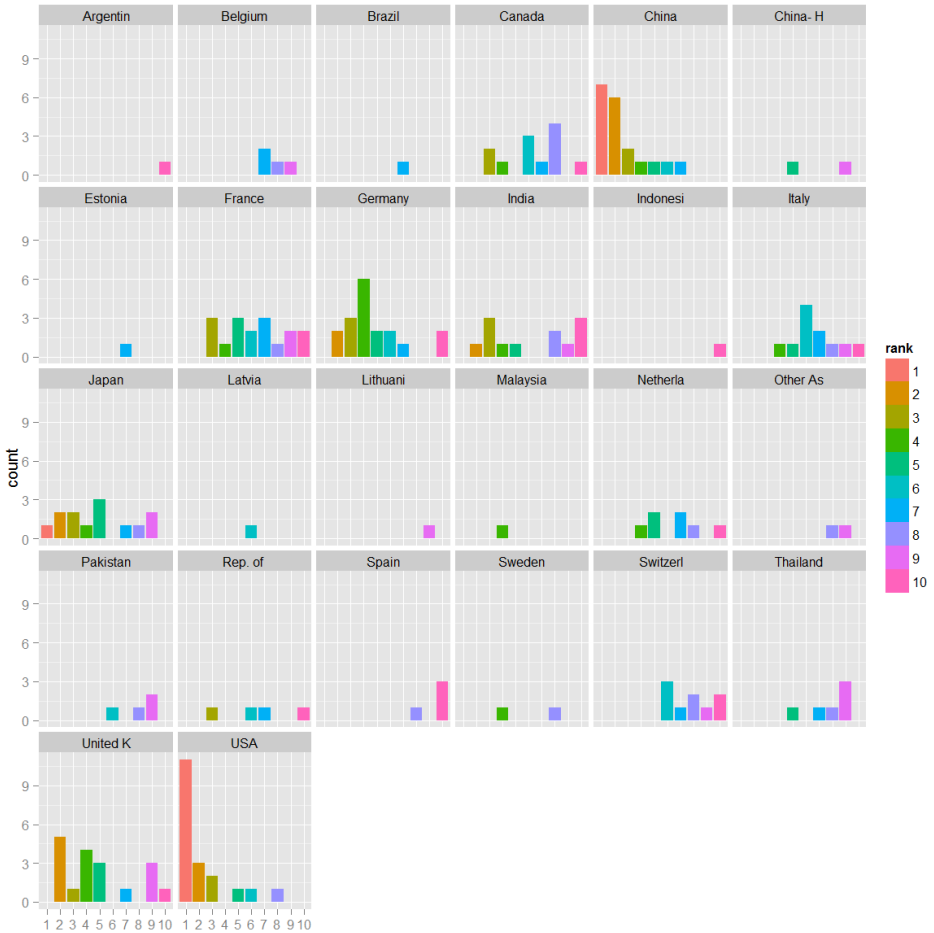




Different commodities

Outputs – Network Analysis

Countries that were counted at least one time in the ten best ranks after application on the PageRank algorithm on each network for 2012



Findings

Relevance
Comprehensive analysis of global value chain through trade networks in all economic sectors is crucial part to better understand international trade and new approaches as those we experimented are needed

Technology
8 different tools/languages were used to work with the data

Starting from data in basic text format made easy to switch from one tool to another

Technology
Processing data with the Sandbox provided evident advantages in terms of processing time and manageable size wrt current tool used at UNSD (Relational DB)

Findings

Methodology
The methodology used to prepare and analyse trade data is not new but the sandbox environment enables comprehensive analysis of the whole data set

Methodology
Novel methods for "automatic" detection of network clusters through machine learning approaches are on the agenda and should be tested before the end of the project

How far are we from something that can be published?

- The objective was not to produce statistics. However analysis such as those produced in this experiment are normally published by international research centers and institutions.

What case has been made for the future of the sandbox?

- It is possible to use big data tools and technologies (Hadoop, Pig, Hive, Spark and Gephi) in processing and analyzing large volume of trade data
- The easiness of setting up the data environment, powerful computing power and availability of built-in libraries to analyse networks may change the way trade analysts work

INTRODUCTION

ACTIVITIES

FUTURE OF SANDBOX

OUTCOMES

Beyond the Big Data Project

- Extend access to the Sandbox beyond the current project (December 2015)

  Based on strong interest from a number of statistical organisations
- ICHEC is willing to continue to provide the Sandbox as a service to the international statistical community, on a non-profit basis
- Users will be required to pay an annual subscription to cover the costs of technical support, hardware upgrades and installation of software

below: some use cases for future Sandbox …

The Sandbox: Use Cases

Running experiments and pilots

The sandbox can be used for experiments involving creating and evaluating **new software programmes**, developing **new methodologies** and exploring the potential of **new data sources**

This use case extends the current role of the sandbox beyond Big Data, and encompasses **all types of data sources**

The Sandbox: Use Cases

Testing

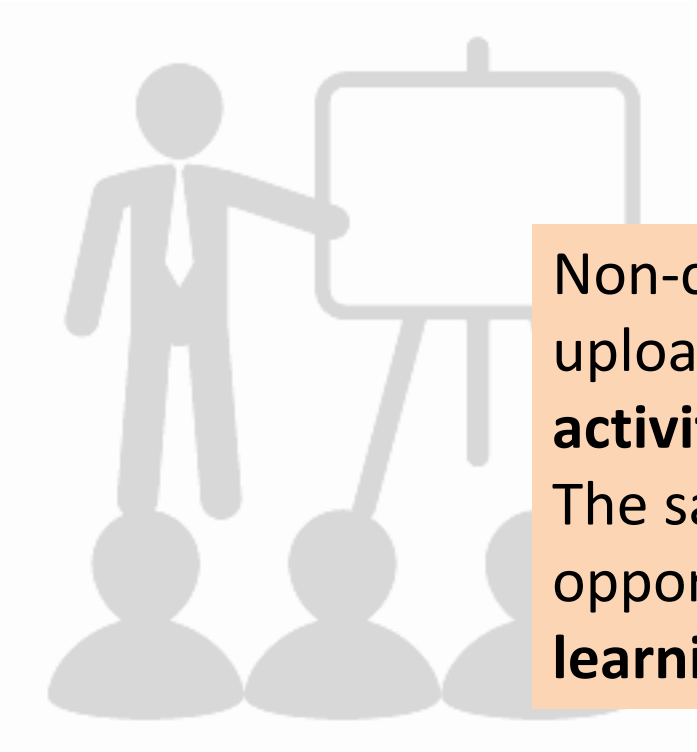
Setting up and testing of statistical pre-production processes is also possible in the Sandbox, including simulating complete workflows and process interactions

The environment could be used also for testing other kind of software, beyond Big Data tools

The Sandbox: Use Cases

Training

The sandbox can be used as a platform for supporting training courses. It can run special software for high performance computing which **cannot be installed or run on standard computers**



Non-confidential demonstration datasets can be uploaded and shared, facilitating **shared training activities** across organisations

The sandbox environment also allows statisticians opportunities for self-learning, e-learning and **learning by doing**

The Sandbox: Use Cases

Supporting the implementation of the Common Statistical Production Architecture (CSPA)

The sandbox can be used as a statistical laboratory where researchers can **jointly develop and test** new CSPA-compliant software

The Sandbox: Use Cases

Data Hub

The sandbox also provides a **shared data repository** (subject to confidentiality constraints)

It can be used to **share non-confidential data sets** that cover multiple countries, as well as public-use micro-data sets

The Sandbox: Use Cases

Big Data Technologies for Statistics

Big Data technologies have proven to be usable for many of the elaborations standards of statistical data

Parallel processing and NoSQL seem to be able to effectively backup the traditional software like RDBMS and old file systems

What you pay

Subscription fee **10k€ per year**

What you get

- Access to the Sandbox, shared tools, datasets and other resources for your staff
- Access to international collaboration projects and opportunities
- Technical support to keep the Sandbox infrastructure up to date and relevant to your needs
- Support for coordination of experiments between countries and for knowledge sharing

How to Subscribe

- Expression of interest coordinated by UNECE support.stat@unece.org
- Each subscriber will have a seat on the Strategic Advisory Board, a new group which will oversee Sandbox operations
  - collectively decide, in consultation with ICHEC, on subscription levels and priorities for expenditure on software and hardware
- Any organisation producing official statistics can subscribe to the Sandbox
  - Other organisations may be considered on a case-by-case basis subject to the approval of the Strategic Advisory Board
- Because the Sandbox activities are closely linked to the work of the HLG-MOS, the UNECE has been asked to facilitate contacts between the official statistical community and ICHEC, and support the functioning of the Strategic Advisory Board

INTRODUCTION

ACTIVITIES

FUTURE OF SANDBOX

OUTCOMES

Summary of Activities Outcomes

	Publishable products	Sandbox value
	Yes, as «alternative» statistics	Storage and collection Use of tools
	Already published in some countries	Sharing of methods
	Further work required	Sharing of methods Use of tools
	Yes, as research	Storage and collection Use of tools

Technology

- Proved use of big data technology to process "traditional" statistical data more efficiently
  - Novel trend in statistical organizations
- Proved practical advantage in using Sandbox tools over "traditional" ones also when dealing with "medium"-size data
- Different tools could be used in the same experiment on the same dataset

Use of Tools

Hive				✓
Pig	✓			✓
RHadoop	✓			✓
Elasticsearch		✓	✓	
Spark				✓
Python	✓		✓	✓
R	✓	✓		✓
Graph visualization				✓

Sharing

- Consolidated the importance and the advantages of working in a cooperative way
- Shared knowledge on tools, methods and solutions
- Gathered lessons learned from activity made also outside the project
- Great value from Sandbox approach
  - Common environment ready in zero time, with no need for installations, configurations etc.

“Usable” outcomes

- Datasets
  - wikistat, comtrade, tweets
- Tools
  - web scraping, twitter collection, wikistats preprocessing
- Training material on tools and sources

Sources

- Difficult to find “quality” sources
- Public data is limited in terms of expected quality and/or requires a lot of processing
- Privacy issues always present even in apparently public sources (e.g. limitations in the use and sharing of scraped data)
- Two kinds of actions needed
  - negotiations and agreements with providers
  - political actions at legislative level

Big Data Features

- Statistics based on big data sources will be *different* from what we have today
  - new sources can cover aspects of reality that are not covered by traditional ones
- Methods should adapt to this
  - accept different definition of quality
  - wider interpretation of results (e.g. consider distorsions due to events)
- We should learn to accept inherent instability of sources in short-long term

Value of the Sandbox

- The Sandbox represents **the fastest route** available to statistical organizations **for starting with Big Data**
- It offers several features that facilitate the approach to Big Data and data science
  - An infrastructure for big data processing ready for being used at a low subscription cost
  - Software already installed and proved/tested
  - Shared datasets instantly available
  - Tools and material for capacity building
- It is driven by the community
  - Worldwide network of organizations, with a catalogue of initiatives
  - A place for collaboration on methods and products, not only related to Big Data
  - A unique container of experience in management of statistical data

Concluding Remarks

- **Huge** thanks to all participants and to the HLG
- We feel we made great steps ahead from where we started two years ago
- Our hope is to not “waste” the experience
  - tools, methods, knowledge, networking
  - connect the Sandbox with national experiences and with “other sandboxes”

THANK YOU!