INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Utilização de fontes *Big Data* para a produção das Estatísticas Oficiais

25ª Reunião Plenária do CSE - Lisboa, 02 de Julho de 2018

BIG DATA

Lots of fish at sea

Which can be used!!!

Questions?

Where to look?

How to Collect?

How to Process?

How to Store?

How to Analyze?

Data Driven Models?

How to Keep Efficiency?

PROJECTS

ESSnet on Big Data

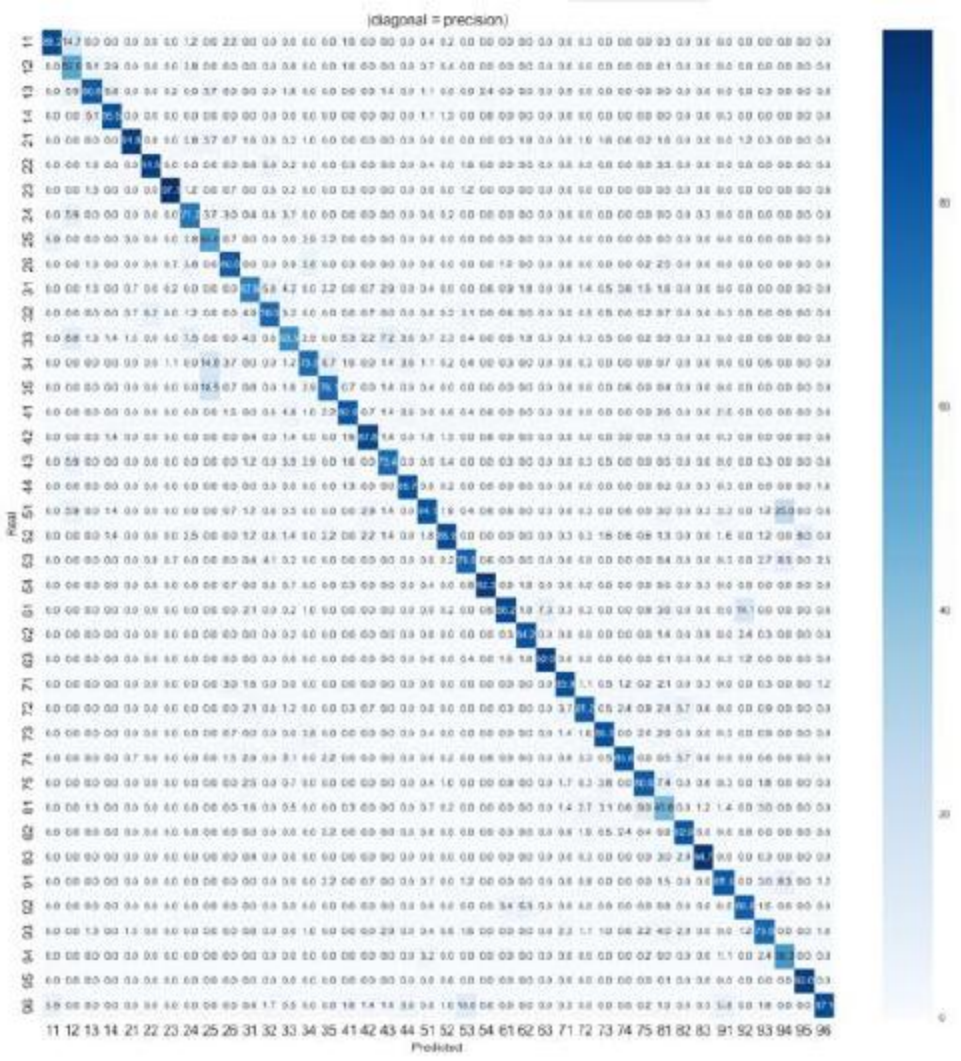
Phase one - Jan/17 to May/18

- From January 2017 to May 2018
- Involved in several work packages
- Using different sources/methods and approaches

Web Scrapping / Job Vacancies

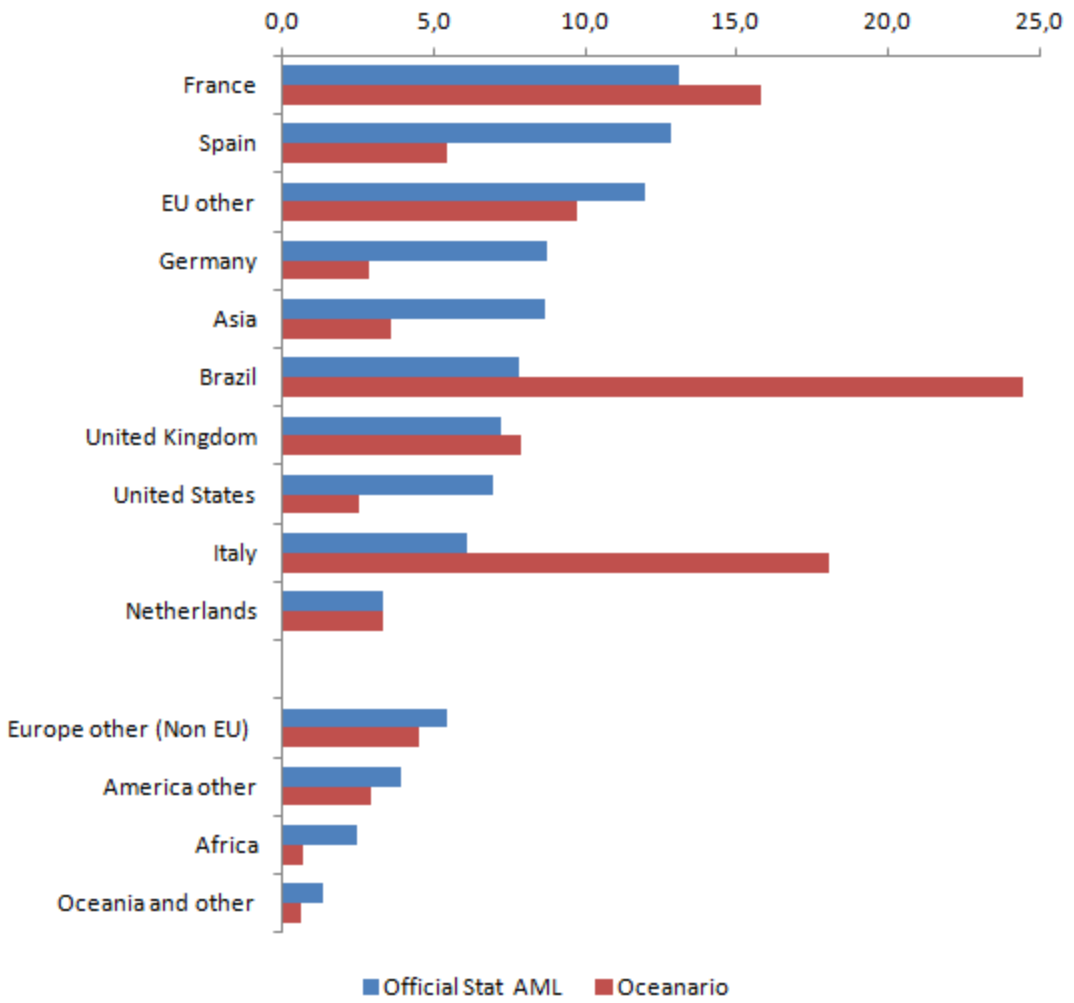
- Web scrapping
- Text analysis and classification
- Machine Learning – SVM with Linear Kernel

- Techniques usable for classification problems not only on Big Data



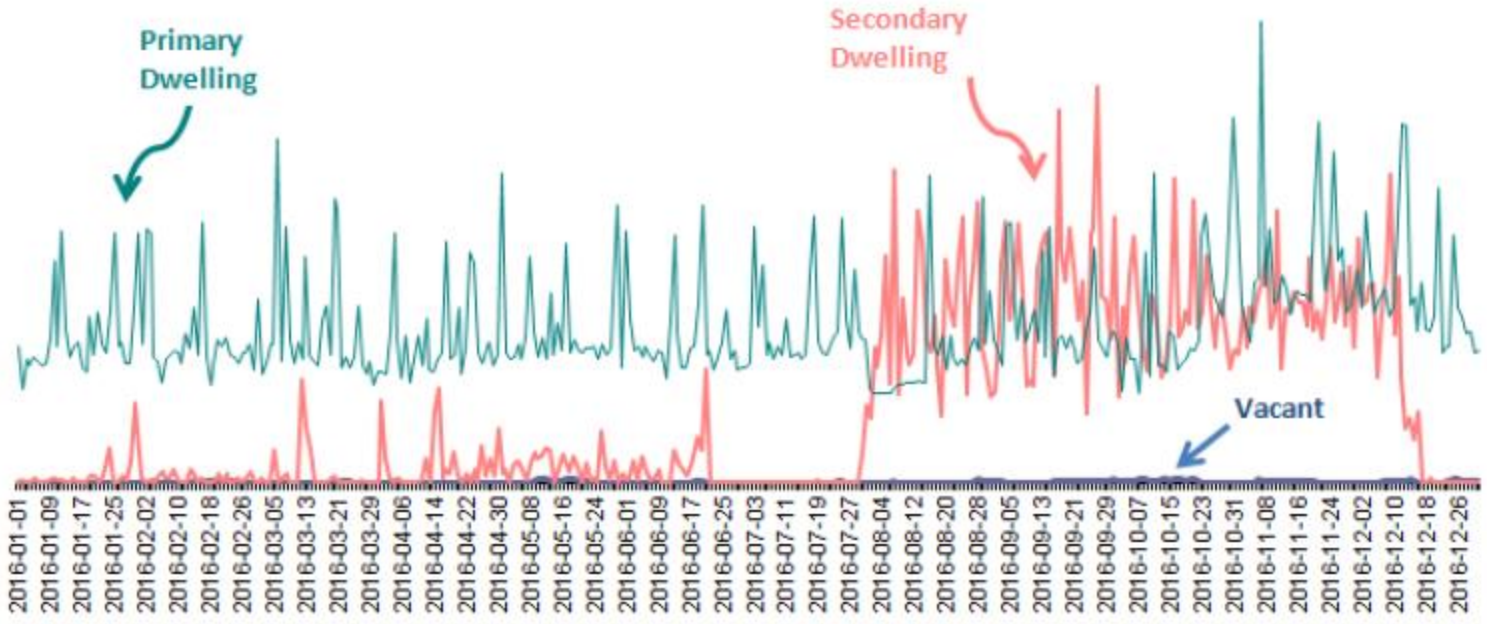
Web Scrapping / Tourism

- Web scrapping for content –static and dynamic
- Processing – clean and normalize
- Classification Tasks
- Language review
- Data Analysis - for example why is Brasil over represented?



Smart meters / Electricity Consumption

- Use electricity consumption to classify dwellings as primary, secondary or vacant
- Clustering techniques used
- Aggregated data Linkage by postal code



SAR / Economic Indicators

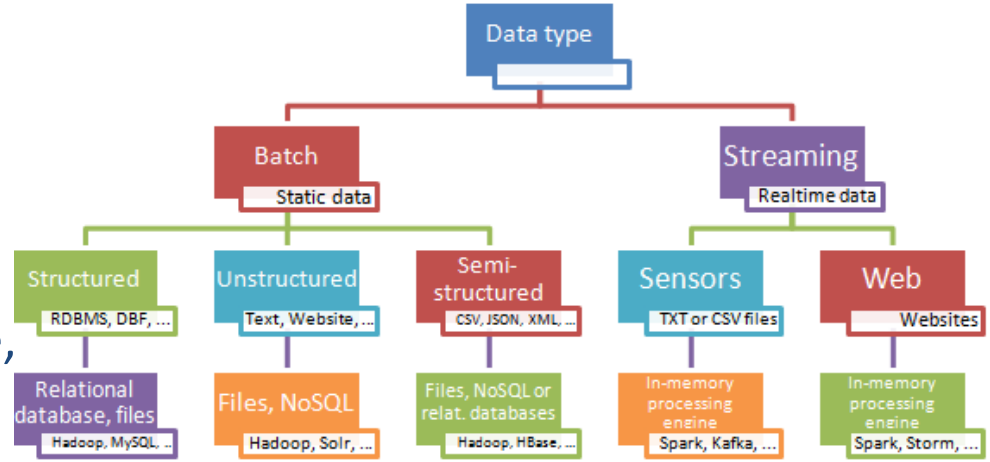
- Volume of exports (INE, 2018), and overnight stays, both collected at NUTS III level as predictors
- Use time series with quarterly data from 2007 to 2017, for 23 different regions, where with each time series having 40 time periods
- SAR (Spatial Autoregressive) is used

Results of the models

Models	Components	Equation	Results
Simple Linear Regression		$Y = \beta_0 + \beta_1 * Nights + factor(região) + \varepsilon$	$R^2 = 0.933$
SAR (Spatial autoregressive model)	Region	$Y = \rho W_Y + X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$	$R^2 = 0.941$ $AIC = 154772.1$
	Region and time	$Y = \rho W_Y + X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$	$R^2 = 0.953$ $AIC = 154119.4$
SEM (spatial error model)	Region	$Y = X\beta + u$ $u = \lambda W_u + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$	$R^2 = 0.936$
	Region and time	$Y = X\beta + u$ $u = \lambda W_u + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$	$R^2 = 0.952$

Methodology

- Methodology Issues:
  - Coverage, accuracy, etc
- Quality Issues:
  - Measurement error, Linkability, etc
- IT issues:
  - Data source integration, processing, big data life cycle, fit to GSBPM



- Crosscut issues to all projects that must include big data for official statistics production

Big Data Hackathon 2017

Subject: Help tackling Skills mismatch in EU

- "Skills development are essential in the emerging new economy and fast-changing labour market"
- "Qualification and skill mismatches entail significant economic and social costs for individuals and firms"

- **Data:** Eurostat datasets; webscrapped data from job portals
- **Aim:** help policy makers, job providers and/or jobseekers
- **Organization:** Eurostat and CEDEFOP
- 5th place in 22 EU teams (13th-15th March 2017, Brussels)

Big Data Hackathon 2017

Project:

- Develop concept of Labour Market Attractiveness.
- Comprehensive and scalable framework to: visualize labour market datasets; cluster EU regions; construct Eigenvariables; and establish associations between relevant indicators and characteristics of the labour market.

Side-results:

- Presented in 3 academic conferences
- 1 Statistical Working Paper and 1 Conference Proceeding
- https://github.com/jsollari/EUhackathon2017
- http://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-TC-18-002

ESSnet on Big Data – Phase Two

Phase Two – Nov/18 to Nov/20

- *submission date 20th of September*
- From November 2018 to November 2020
- Possibility of Involvement in several areas
- Invited to lead a Work Package

Phase Two – Nov/18 to Nov/20

Work Packages will cover:

1. Implementation
2. Research and Development
3. Study and Exploration

Phase Two – Nov/18 to Nov/20

The call being prepared

Description	Countries interested	
	Number	Designation
Use of financial transactions data	4 to 5	BG, (IT), NO, PL, **PT**
Remote sensing	5 to 6	DE, FR, (IT), PL, **PT**, SK
Mobile network operator data	13	...
Innovative sources and methods for tourism statistics	8	AT, DE, DK, GR, IE, PL, **PT**, SK

Phase Two – Nov/18 to Nov/20

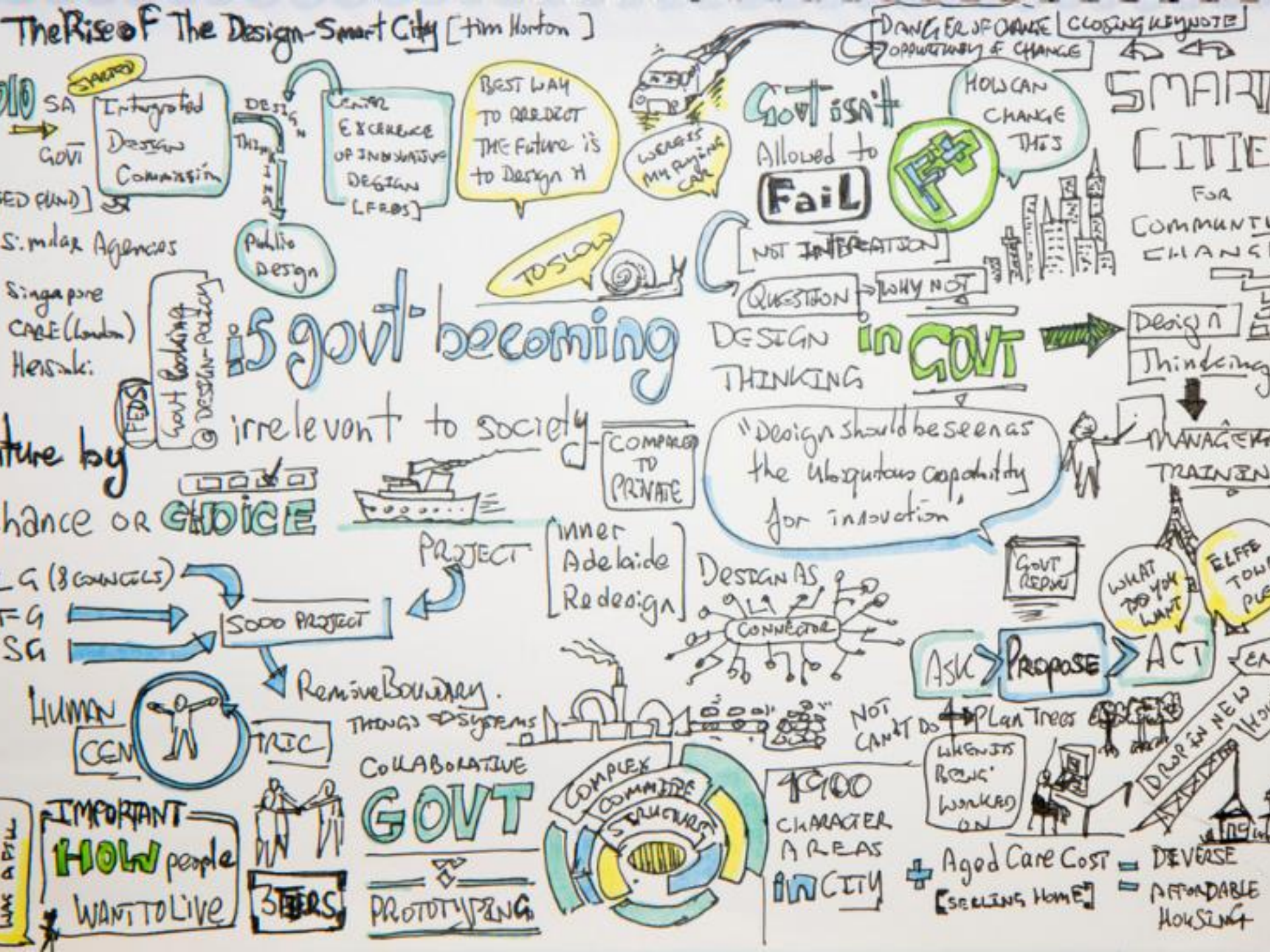
The call being prepared

Description	Countries interested	
	Number	Designation
Broader View	8	IT, BG, SE, FR, DK, NO, PL, **PT**
InDepth view - smart farming	3	AT, PL, **PT**

Call to FCT on Modernization

Challenges

BigMove


The Rise of The Design-Smart City [tim Horton]
DANGER OF CHANGE
OPPORTUNITY & CHANGE
CLOSING KEYNOTE
SMART CITIES FOR COMMUNITY CHANGE
GOVT isn't Allowed to FAIL
HOW CAN CHANGE THIS
BEST WAY TO PREDICT THE FUTURE IS TO DESIGN IT
WHERE'S MY FLYING CAR
Integrated Design Commission
Centre Excellence of Innovative Design [CEDIS]
Public Design
Similar Agencies
Singapore
CABE (London)
Helsinki
FEDS
Govt Looking @ Design-Policy
is govt becoming irrelevant to society
TOO SLOW
NOT INTERACTION
QUESTION
WHY NOT
DESIGN THINKING in GOVT
Design Thinking
MANAGERS TRAINING
"Design should be seen as the ubiquitous capability for innovation"
COMPARED TO PRIVATE
ture by
hance or CHOICE
PROJECT
Inner Adelaide Redesign
Design As
CONNECTOR
GOVT REDESIGN
WHAT DO YOU WANT
EIFFEL TOWER
ASK
PROPOSE
ACT
G (3 COUNCILS)
SOOO PROJECT
Remove Boundary.
THINGS SYSTEMS
NOT CAN'T DO
Plan Trees
HUMAN CENTRIC
COLLABORATIVE
GOVT
PROTOTYPING
COMPLEX COMMUNITY STRUCTURES
1900 CHARACTER AREAS in CITY
WHEN IT'S REALLY WORKED ON
DROP IN NEW
Aged Care Cost = DIVERSE AFFORDABLE HOUSING
[SELLING HOME]
IMPORTANT HOW people WANT TO LIVE
3 TIERS

BigMove

Project information:

- 36 months duration
- Use Big Data – primarily from sensors
- Focus on the city of Lisbon
- Multidisciplinary team from INE and university – 5 internal researchers

BigMove

Project statistical aims:

- Knowing how people move in Lisbon
- Identify the patterns of commuting, longer and shorter movements to assess the quality of life in the city
- Realize the intensity of traffic and transport
- Predict and anticipate how events disturb the city

BigMove

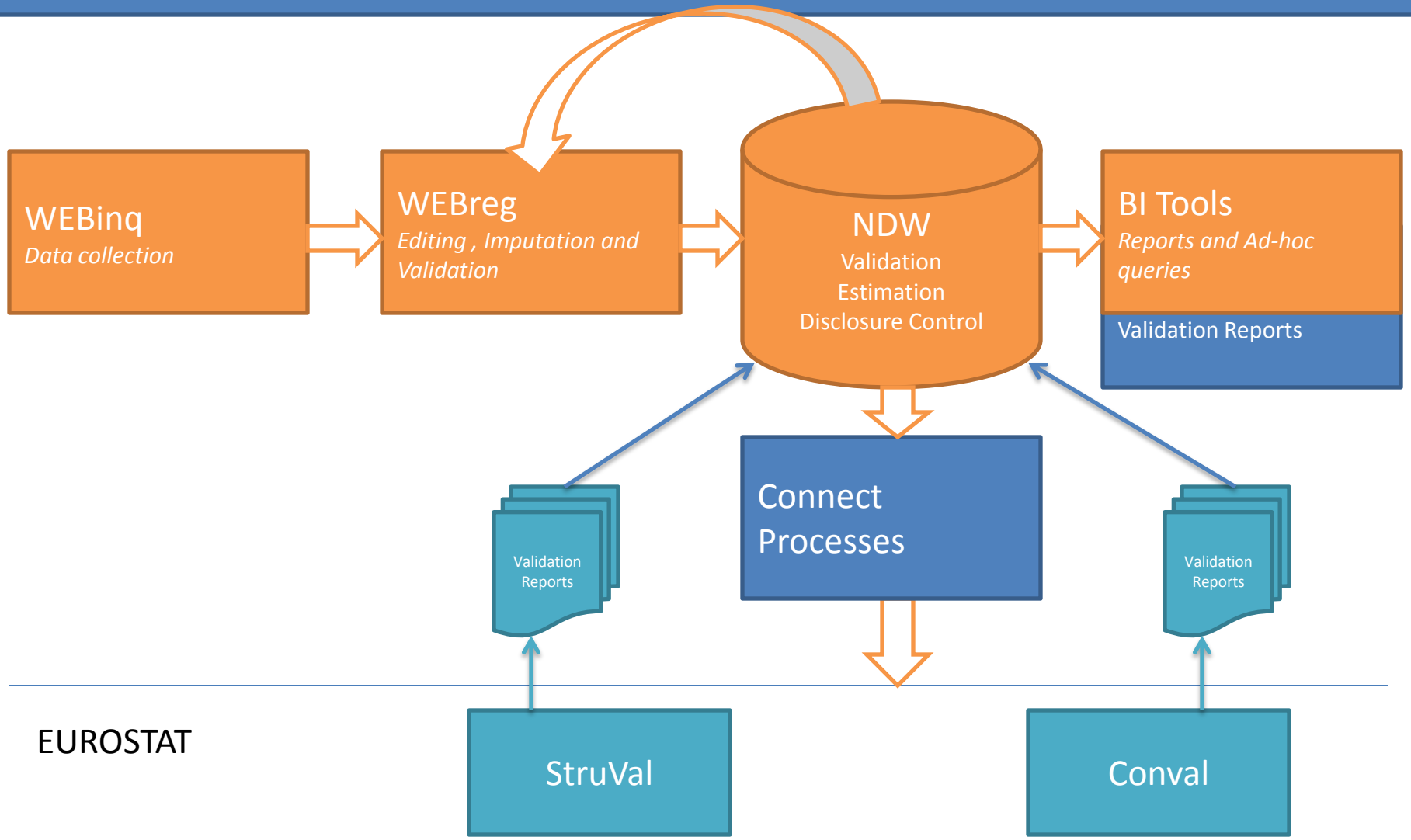
Project generic aims:

- Endow INE with a structure capable of dealing with big data in terms of storage and processing
- Enable INE's human resources to use data analytics and big data
- Include data discovery, machine learning and deep science techniques in the processing methods available at INE
- Gain new ways of visualizing data appropriate to the challenges that big data poses to INE

ESSnet on Validation

Validation

Situation at INE



Validation – Jan/17 to Feb/18

- From January 2017 to February 2018
- Leadership of the Work Package on Implementation
- Leadership of the Work Package on accessing the cost benefit analyses of the validation tools
- Development of multi-criteria tool for cost/benefit analyses

Validation – Jan/17 to Feb/18

Covered the scenarios for data validation of the Business Architecture for the ESS

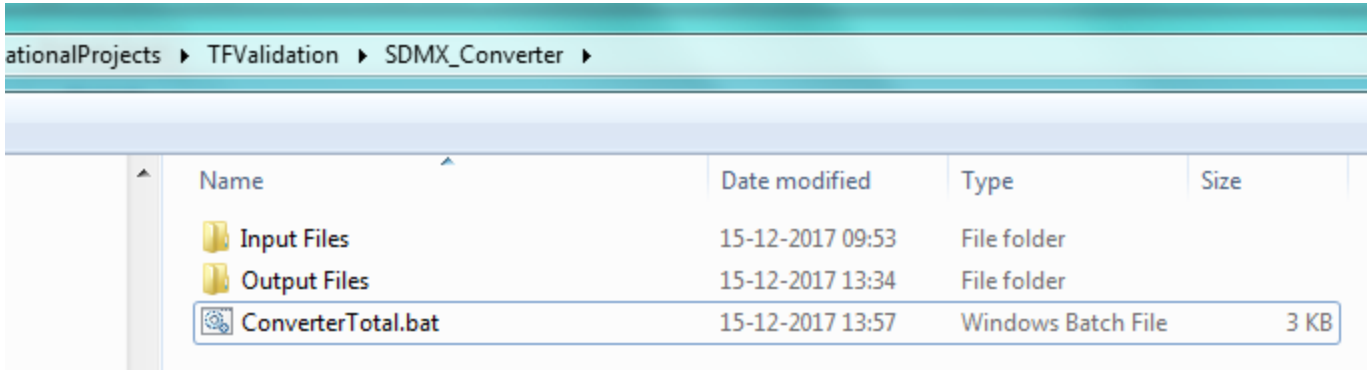
- Structural Validation
- Content Validation
- EDAMIS

Validation – Jan/17 to Feb/18

Results:

- Assessments on Structural Validation
- Assessments on Content Validation
- Assessments on VTL language
- Pilot Automation of the Structural and Content Validation for the National Accounts

We've automated the process




```
36  set data=%(                                                              %
37  for %%F in
38     rem set file=%%~nxF
39     set file=%%~nF
40     rem echo "FILE" %file%
41     set extensao=%%~xF
42     set id=!file:~0,14!
43     rem echo "FILE>"  !file!
44     rem echo "... ID >" !id!
45     rem echo "....extensao >"  !extensao!
46     rem COPIAR o template para Header
47     COPY %CLASSPATHInput%\Template.prop %CLASSPATHInput%\Header.prop
48     echo.>> %CLASSPATHInput%\Header.prop
49     echo header.id=!id! >> %CLASSPATHInput%\Header.prop
50     echo header.prepared=!data! >> %CLASSPATHInput%\Header.prop
51     Call %CLASSPATH%\Converter.bat -header_file %CLASSPATHInput%\Header.prop -dsd_file %CLASSPATHInput%\DSD.
52     set /a nfiles+=1
53  )
54  cls
```

Validation – Jan/19 to Dec/19

- From January 2019 to December 2019
- Single grant agreement
- Extending validation of the outputs to other domains: Agriculture
- Compare validation rules in VTL and SQL to achieve a template with the core rules that cover 80% of the rules adopted by Eurostat

Centre of Excellence on SDW

Data Warehouse

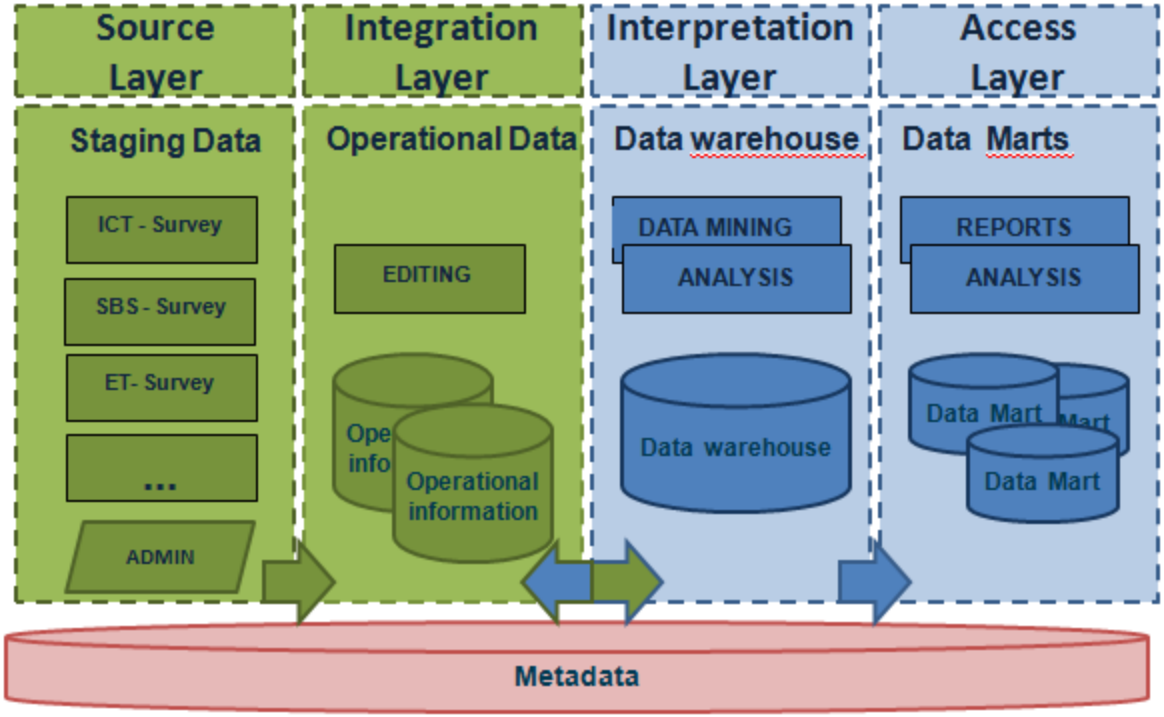
CoE on Statistical Data warehouse

- From November 2016 to November 2018
- IV mandate in the Centre of Excellence on Data Warehouse, after 3 ESSnets (since2011)
- Internships since 2015 in the data warehouse area recognizing our Best Practices

CoE on Statistical Data warehouse

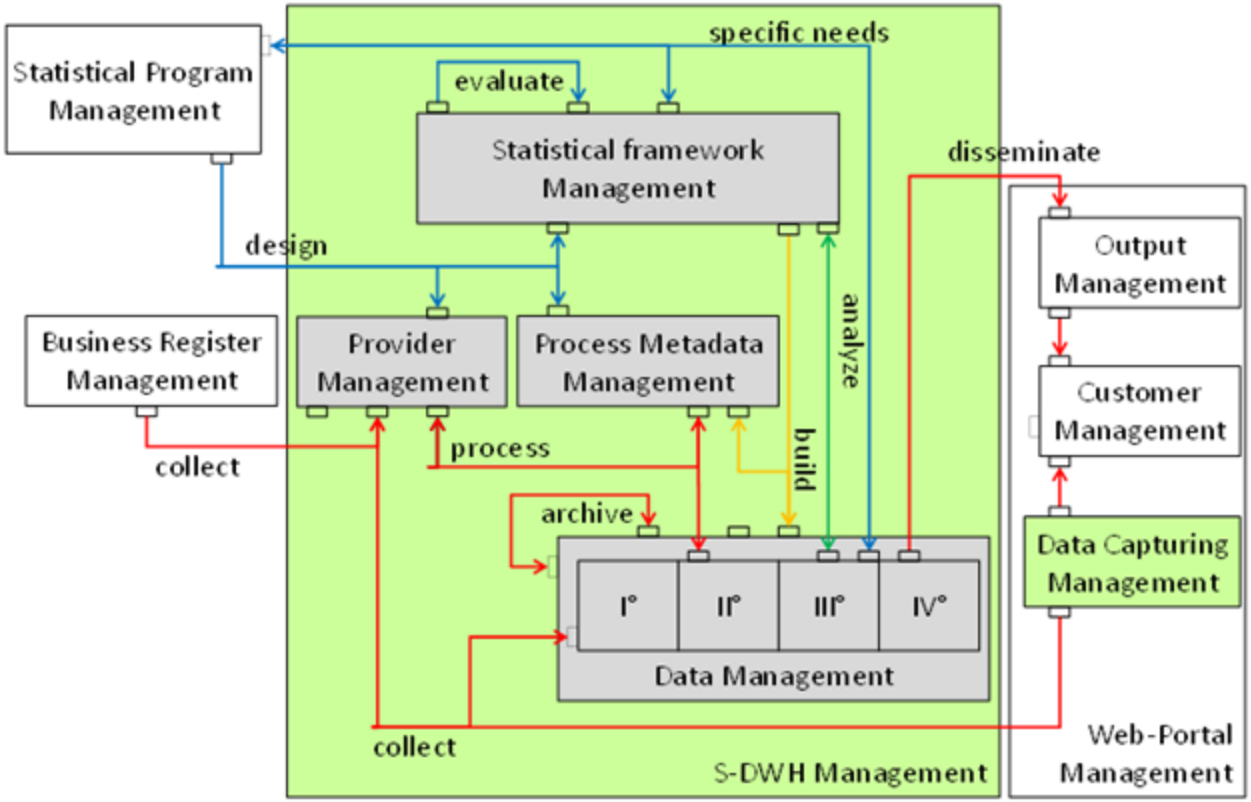
Lines of Work:

- Documenting of Best practices in the ESS countries

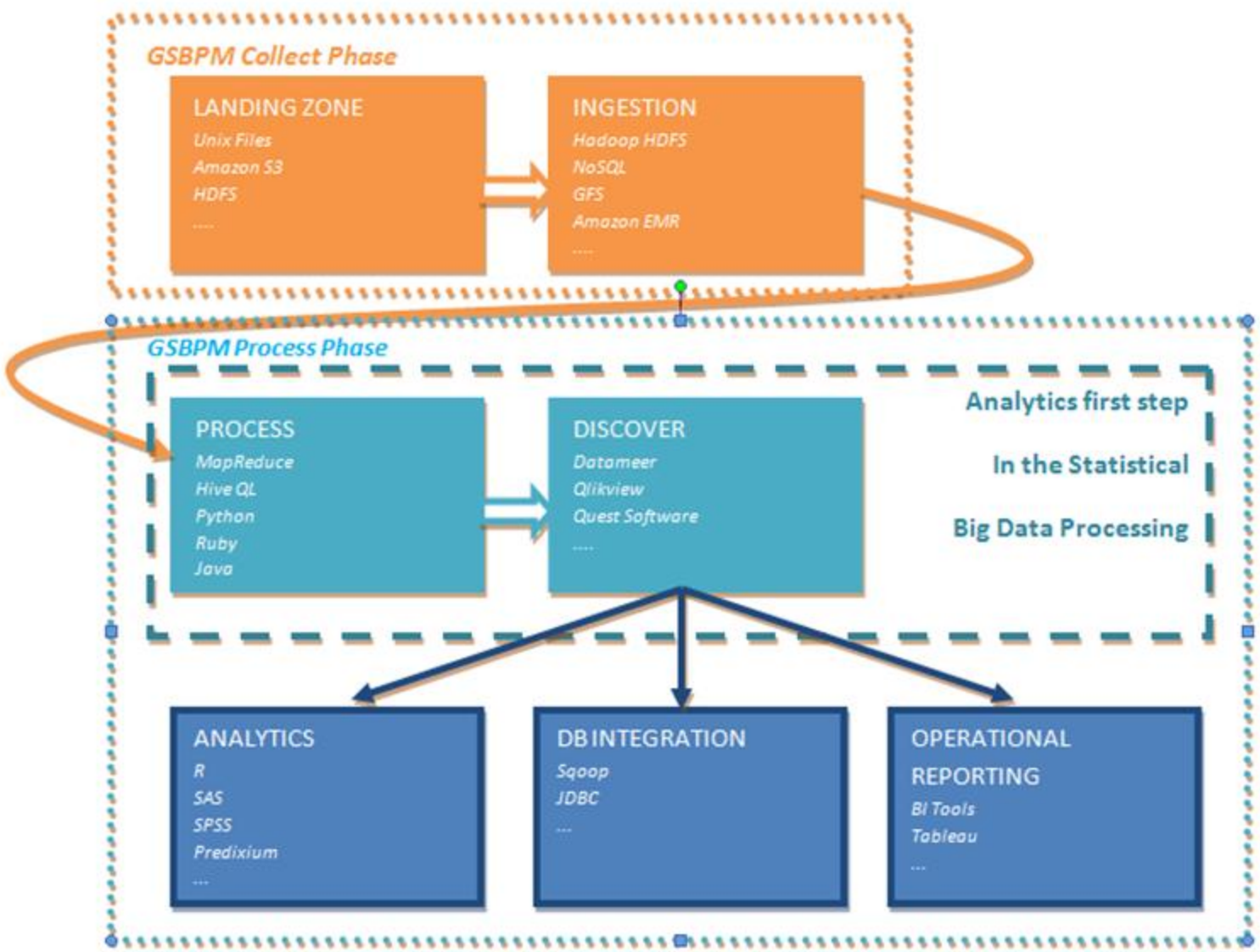


CoE on Statistical Data warehouse

Preparing a Handbook for implementing a Statistical Data Warehouse



Big Data Conceptual Platform



ESSnet on Services

ESSnet on Services

ESSnet on Services

- Call submitted on the 27th June
- Duration 24 months
- Focus on:
  - Services development and reuse
  - Understand the barriers to services reuse
  - Build on success stories to promote services reuse across the ESS

ESSnet on Services – Dissemination WP

Task/Work package number	4	Start Month: End Month:	T+1 T+24
Title	Newsletter		
Partner	**1 (bold for leader)**	2	3	4	5	6	7	8	9	10	11	12	13	
Objectives Increase the awareness in the countries of the ESS to the results of this ESSnet														
Description of work During the intermediate months create a newsletter following the major developments, drawbacks an sucesses of the services inside this essnet. **Sub-Tasks:** • Investigate every work package in the essnet with a differente perspective – creating a narrative fo each • Follow up on particular details and facts • Prepare and write the newsletters														
Deliverables Newsletters														

Task/Work package number	4	Start Month: End Month:	T+1 T+18
Title	Experts meeting		
Partner	**1 (bold for leader)**	2	3	4	5	6	7	8	9	10	11	12	13	
Objectives Bridge the gap between Service developers and re-users														
Description of work Investigate why there are more candidates to services development than to service re-use. **Sub-Tasks:** • Organize "Mind the Gap" meetings • Uncover the barriers to services re-use • Discover ways to overcome those barriers • Collect and build on sucess stories on services re-use														
Deliverables Report on "Mind the Gap" solutions														

Distribution of man power for the duration of the action

ESSnet on Services – Dissemination WP

Task/Work package number	4	Start Month: End Month:	T+1 T+12
Title	Survey		
Partner	**1 (bold for leader)**	2	3	4	5	6	7	8	9	10	11	12	13	
Objectives Know the current status and maturity of the countries in the ESS toward service adoption														
Description of work Elaborate a survey with the aim of knowing the current status and maturity of the countries in the ESS towar service adoption, and the areas and processes where those will be most needed or required. **Sub-Tasks:** • Prepare the survey • Run the Survey • Treat the survey results • Disseminate the survey results														
Deliverables Report on the results of the survey														

Task/Work package number	4	Start Month: End Month:	T+1 T+24
Title	Sinder		
Partner	**1 (bold for leader)**	2	3	4	5	6	7	8	9	10	11	12	13	
Objectives Develop a way to let NSIs discover the Services most suited to their profile														
Description of work Based on the results of the survey cluster typical profiles and develop a binary tree to guide a NSI toward the Services that suit them. **Sub-Tasks:** • Create the typical profiles • Categorize the services • Create a yes/no workflow - Sinder														
Deliverables Report on Sinder and Sinder itself														

Data Analytics

Data Analytics

Data Analytics

- First preparatory meeting May 2018
- First conference September 2018
- Focus on:
  - How to use machine learning for official statistical production
  - What does using Big Data requires from data analytics at the NSIs
  - Can a Logical Statistical Data warehouse (LSDw) assists us on the process?

Trusted Smart Statistics

Trusted Smart Statistics

Trusted Smart Statistics

- First conference April 2018
- Second conference January 2019
- Focus on:
  - How can smart statistics overcome by aggregation of data some legal issues that big data poses?
  - Smart data coming from sensors is less prone to change than a website and so more dependable to produce official statistics
  - Which methods to ensure coverage, accuracy and measurement error should be applied to Smart Statistics to make them Trustable?

Instituto Nacional de Estatística
Departamento de Metodologia e Sistemas de Informação
jorge.magalhaes@ine.pt