



Horizon 2020
European Union Funding
for Research & Innovation

Project 'cities2030' | H2020 ID | 101000640 | 'Co-creating resilient and sustainable food systems towards FOOD2030' | www.cities2030.eu

cities2030

D6.2 Data sources and integration report



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101000640

Deliverable D6.2

Prepared by P21 | Edited by P19, P20, P21, P35, P37 | Checked and reviewed by EX COMM | Approved by P20
Rev 1.2 - June 2022

Document information

Key information	Data
Project reference number	101000640
Project acronym	'cities2030'
Project title	'Co-creating resilient and sustainable food systems towards FOOD2030'
Project start date	October 1 st , 2020
Duration	48 months
PPaCO	Mr Nicola CAMATTI
Project website	cities2030.eu
Work package (WP)	6
WP leader	UPM (P20)
WP co-leader	WIT (21)
Contributors and authors	WIT(P21): Paul Malone, Mohit Taneja, Micheal Crotty UNIL (P35): Marharyta Aleksandrova UPM (P20): Ramón Alcarria, Borja Bordel, Jesús Sánchez López, Diego Martín, Miguel Manso
Peer reviewers	All WP6 partners
Peer review start date	14/10/2022
Peer review end date	7/11/2022
Document type ¹	R
Document/file name	Cities2030_D6.2_Data_Sources_v1.4
Document title	D6.2 – Data sources and integration report
Deliverable number	D6.2
Abstract	The objective of document D6.2 describes the work performed in Task 6.2, which focuses on acquiring, preparing, analysing, fusing and integrating the incoming datasets that support the project use cases in a horizontal manner, as well as provide the mechanisms to access and share them. In handling such large and heterogeneous datasets, techniques such as distributed compression and compressed learning can be taken into consideration for effective data management. Data mining methods such as clustering, and feature extraction will be investigated to prepare data conducting effective analysis. Distributed learning techniques such as Federated learning and representative learning approaches can then be taken into consideration for developing collaborative data integration and analysis mechanisms. Additionally, techniques of prediction with guarantees such as causal and conformal learning will be investigated.
Deliverable date (DoW)	September 30th, 2022
PM approval date/version	14/11/2022 Version v.1.4
Submission date	November 15th, 2022
For public dissemination Yes/NO	YES

¹ R: document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: websites, patents filing, press & media actions, videos, etc.; OTHER: software, technical diagram, etc.

Deliverable D6.2

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

All 'cities2030' consortium members are also committed to publish accurate and up-to-date information and take the greatest care to do so. However, the 'cities2030' consortium members cannot accept liability for any inaccuracies or omissions, nor do they accept liability for any direct, indirect, special, consequential or other losses or damages of any kind arising out of the use of this information.

Copyright notice



This work by Parties of the Cities2030 Consortium will be licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

Acknowledgement



The project 'Co-creating resilient and sustainable food systems towards food2030' (Cities2030) has received funding from the European Union Horizon 2020 programme under grant number 101000640

Citation

Be so kind as to cite this work as:

'Co-creating resilient and sustainable food systems towards food2030', Cities2030 Consortium, Guidelines for a data-driven CRFS management platform – Update Cities2030 consortium under the supervision of the project's coordinator.

Legislation

Legislation H2020 Framework Programme – Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 – The Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p. 104).

Euratom Research and Training Programme (2014-2018) – Council Regulation (Euratom) No 1314/2013 of 16 December 2013 on the Research and Training Programme of the European Atomic Energy Community (2014-2018) complementing the Horizon 2020 – The Framework Programme for Research and Innovation (OJ L 347, 20.12.2013, p. 948).

H2020 Specific Programme – Council Decision 2013/743/EU of 3 December 2013 establishing the Specific Programme Implementing Horizon 2020 – The Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p. 965).

Rules for Participation (RfP) – Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 of December 2013 laying down the rules for the participation and dissemination in Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p.81).

Financial Regulation (FR) – Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council of 25 October 2012 on the financial rules applicable to the general budget of the European Union (OJ L 298, 26.10.2012, p.1).

Rules of Application (RAP) – Commission Regulation (EC, Euratom) No 1268/2012 of 29 October 2012 on the rules of application of I Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council on the financial rules applicable to the general budget of the Union (OJ L 298, 26.10.2012, p.1).

Document history

Version (v)	Date ²	Comment	Author	Status ³
0.1	7/12/2020	Initial structure	UPM	-
0.2	01/02/2021	Introduction first draft	All partners	Drafted
0.3	08/05/2021	Methodology for data integration and analysis	UPM	Drafted
0.4	15/06/2021	Revision of methodology and incorporation of initial component description	WIT	Drafted
1.0	10/11/2021	Adding State of the art section	UPM, WIT	Drafted
1.1	03/03/2022	Architecture provision and revision	UPM	Drafted
1.2	11/08/2022	Component description	WIT, all	Drafted
1.3	14/10/2022	Early final draft	All partners	Drafted
1.4	07/11/2022	Integration of last contributions and final check before PMO validation	WIT	Completed
1.4	14/11/2022	Formal layout check	P02 EPC	Completed

List of Figures

Figure 1. Methodology for the development of the data integration and analysis subsystem	12
Figure 1. Centralised versus Distributed versus Federated Learning	30
Figure 2. Categorization of Federated Learning	31
Figure 4. Traditional Machine Learning versus Conformal Learning	34
Figure 4. Calculation of calibration scores for classification	35
Figure 5. Sample BPMN extension definition	46
Figure 6. SEMEX Portal (semex.io)	47
Figure 7. GEneral Multilingual Environmental Thesaurus	48
Figure 8. SEMEX NLP core system	49

² As per the project's cloud storage if applicable, or per email date if applicable

³ Drafted, completed or validated

Figure 9. Data science process: Adapted from: Blitzsten, J. and Pfister, H., Harvard data science course CS109.....	52
Figure 10. Data integration and analysis subsystem architecture.....	54
Figure 10. Interactions on the Data Integration and Management component.....	55
Figure 12. A schematic representation of the component.....	57
Figure 12. Historical distribution of tweets neutral (NEU), negative (NEG) and positive (POS) tweets	58
Figure 14. Geographical distribution of tweets	58
Figure 14. Results of image analysis	58
Figure 17. Architecture of Geospatial information and services component	59
Figure 18. Geospatial information and services component workflow	60
Figure 19. Code snippet for data collection in Bremerhaven lab	61
Figure 20. Process for obtaining territorial limit for Bremerhaven lab	62
Figure 21. Symbol categorization process in QGIS	62
Figure 22. Result of data curation and preview process	63
Figure 22. Result of data storage process for Seinäjoki lab.....	64
Figure 17. Data Integration, Discovery and Access Micro-services	65
Figure 18. Meta-data structure and inter-relationships.....	66
Figure 18. ARGANDA Legacy Data Infrastructure	97
Figure 19. ARGANDA Use Case proposed solution	97

List of Tables

Table 1. Data integration and analysis components and responsible partners	11
Table 2. Comparison among some Federated Learning implementations	32
Table 3. Learning approaches and challenges	45
Table 4. Optional properties. Geospatial information.....	74
Table 5. Twitter API returns the data in the following format (per tweet)	75
Table 6. Description of the field 'user'	75
Table 7. Description of the field 'place'	76
Table 8. Description of the field 'extended_tweet'	76
Table 9. Properties for real-time data.....	77

Table of Contents

DOCUMENT INFORMATION	1
DISCLAIMER.....	2
COPYRIGHT NOTICE.....	2
ACKNOWLEDGEMENT.....	2
CITATION.....	2
LEGISLATION	2
DOCUMENT HISTORY.....	3
LIST OF FIGURES	3
LIST OF TABLES	4
TABLE OF CONTENTS	4

Deliverable D6.2



GLOSSARY AND ABBREVIATIONS.....	7
---------------------------------	---

1 INTRODUCTION	8
-----------------------------	----------

1.1 SHORT DESCRIPTION OF THE CITIES 2030 PROJECT	9
1.2 SHORT DESCRIPTION OF THE WP6 PACKAGE	9
1.3 PURPOSE OF D6.2 REPORT	9
1.4 RELATION OF THE D6.2 REPORT WITH OTHER WPs AND DELIVERABLES	9
1.5 D6.2 STRUCTURE	10
1.6 ROLES AND RESPONSIBILITIES.....	11

2 METHODOLOGY FOR DATA INTEGRATION AND ANALYSIS	11
--	-----------

2.1 FIRST PHASE	11
2.2 SECOND PHASE.....	11
2.3 THIRD PHASE	12

3 STATE OF THE ART	12
---------------------------------	-----------

3.1 INTRODUCTION	12
------------------------	----

3.1 DATA MINING TASKS, OPERATIONS AND TECHNIQUES.....	14
---	----

3.1.1 Data Warehousing.....	14
3.1.2 Data Cleaning.....	14
3.1.3 Statistical Techniques.....	14
3.1.4 Data Visualization	15
3.1.5 Outlier Detection.....	15
3.1.6 Association	15
3.1.7 Regression.....	16
3.1.8 Classification	16
3.1.9 Clustering.....	17
3.1.10 Predictive Modelling.....	17
3.1.11 Machine Learning.....	18
3.1.12 Neural Networks and Deep Learning	18
3.1.13 Text mining.....	19
3.1.14 Web Mining.....	20

3.2 DATA MINING TOOLS AND TECHNOLOGIES.....	20
---	----

3.2.1 Data Warehousing Technologies	20
---	----

3.2.1.1 PostgreSQL.....	21
3.2.1.2 MariaDB	21
3.2.1.3 MongoDB.....	21
3.2.1.4 Cassandra	21
3.2.1.5 Apache Universe	21
3.2.1.5.1 Hadoop.....	22
3.2.1.5.2 Spark.....	22
3.2.1.5.3 Apache Hive.....	23
3.2.1.5.4 Apache Spatial Information System (SIS)	23
3.2.1.5.5 Other Apache Tools and Technologies	23

3.2.2 Python Scientific Stack	24
-------------------------------------	----

3.2.2.1 Anaconda.....	24
3.2.2.2 NumPy	24
3.2.2.3 Pandas	24
3.2.2.4 Matplotlib	25
3.2.2.5 Statsmodels	25
3.2.2.6 Seaborn.....	25

3.2.3 Dashboard Creation – Business Intelligence (BI) Technologies.....	25
---	----

3.2.3.1 Power BI.....	25
3.2.3.2 Tableau	26
3.2.3.3 Pentaho	26

3.2.4 Machine Learning (ML) and Deep Learning (DL) Frameworks.....	26
--	----

Deliverable D6.2



3.2.4.1	Scikit-learn	27
3.2.4.2	TensorFlow and Keras.....	27
3.2.4.3	PyTorch.....	27
3.2.5	<i>Tools and Technologies Specific to Certain Data Mining Operation</i>	27
3.2.5.1	Gephi	27
3.2.5.2	NLTK.....	28
3.2.5.3	Semex tool (PoliRural project)	28
3.2.6	<i>Cloud-Oriented Solutions</i>	28
3.2.6.1	Amazon Web Services	28
3.2.6.2	Microsoft Azure	29
3.2.6.3	Google Cloud	29
3.3	LEARNING MODELS AND APPROACHES	29
3.3.1	<i>Challenges</i>	29
3.3.2	<i>Distributed Learning.....</i>	30
3.3.3	<i>Federated Learning</i>	31
3.3.3.1	Data partitioning.....	31
3.3.3.2	Aggregation communication	32
3.3.3.3	Implementation challenges	32
3.3.3.3.1	Communication Bottlenecks	33
3.3.3.3.2	Systems heterogeneity.....	33
3.3.3.3.3	Trade-off between privacy and efficiency	33
3.3.4	<i>Conformal learning</i>	33
3.3.4.1	Estimation of calibration scores	34
3.3.4.2	Classification.....	35
3.3.4.3	Regression	35
3.2.6.1	Measuring efficiency and Impact of nonconformity functions.....	36
3.3.4.4	Classification.....	36
3.3.4.5	Regression	38
3.3.4.6	Applications of Conformal Learning	38
3.3.5	<i>Causal analysis</i>	39
3.3.6	<i>Distributed Compression & Compressed Learning</i>	42
3.3.6.1	Sparsification	42
3.3.6.2	Quantisation	43
3.3.6.3	Compressed Learning	43
3.3.7	<i>Summary.....</i>	44
3.4	PREVIOUS PROJECTS.....	44
3.4.1	<i>BUSPrime</i>	44
3.4.1.1	Integration with BPNM	45
3.4.1.2	Extending BPMN	45
3.4.1.3	Analysis of the approach	46
3.4.2	<i>Polirural: SEMEX.....</i>	46
3.4.3	<i>DEMETER.....</i>	49
3.4.3.1	Scope of the data-sources used in DEMETER pilots.....	50
3.4.3.2	Common Data Models and Semantic Interoperability Mechanisms	50
3.4.3.3	Analysis of the approach	51
4	DATA INTEGRATION AND ANALYSIS SUBSYSTEM: ARCHITECTURE AND DESIGN	52
4.1	INTRODUCTION TO DATA INTEGRATION AND ANALYSIS.....	52
4.1.1	<i>A generic data processing solution</i>	53
4.1.1	<i>Pipelines for CRFS.....</i>	53
4.2	DATA INTEGRATION AND MANAGEMENT SUBSYSTEM: ARCHITECTURE AND COMPONENTS	53
4.2.1	<i>Integration and management service</i>	54
4.2.2	<i>Geospatial information</i>	55
4.2.3	<i>Open access repository</i>	55
4.2.4	<i>Social Network and Sentiment analysis component</i>	56
4.2.5	<i>Data mining tool</i>	56
5	SOCIAL NETWORK AND SENTIMENT ANALYSIS COMPONENT	56

6	GEOSPATIAL INFORMATION	59
6.1	ARCHITECTURE AND PROCESS	59
6.2	DATA COLLECTION	61
6.3	DATA CURATION AND PREVIEW	61
6.4	DATA STORAGE	63
6.5	DISSEMINATION AND ENGAGEMENT	64
7	DATA INTEGRATION AND MANAGEMENT SERVICE	64
7.1	MICRO-SERVICE DESCRIPTION	65
7.1.1	<i>Front-end micro-service</i>	65
7.1.1.1	Implementation	65
7.1.2	<i>Search micro-service</i>	65
7.1.2.1	Implementation	65
7.1.3	<i>Metadata micro-service</i>	65
7.1.3.1	Implementation	66
7.1.4	<i>Data view micro-service</i>	66
7.1.4.1	Implementation	66
7.1.5	<i>Datasource ingest micro-service</i>	67
7.1.5.1	Implementation	67
7.2	INFORMATIONAL ARCHITECTURE	67
7.2.1	<i>Meta-data from the Data producer</i>	68
7.2.2	<i>Meta-data automatically assigned</i>	68
7.2.3	<i>Meta-data technical aspects</i>	68
7.3	DATA SOURCES	69
7.3.1	<i>Surveys (Quart de Poblet)</i>	69
7.3.2	<i>Geospatial information</i>	72
7.3.2.1	Data source detailed description	72
7.3.3	<i>Social network information (Twitter)</i>	73
7.3.3.1	Data source detailed description	74
7.3.4	<i>Sensor real-time data (Arganda)</i>	75
7.3.4.1	Data source detailed description	76
7.4	GENERAL ANALYSIS OF RELEVANT EXTERNAL AND INTERNAL DATA SOURCES	77
8	CONCLUSIONS AND NEXT STEPS	78
	ANNEX I - QUART DE POBLET USE CASE	79
	<i>Resources:</i>	79
	<i>Survey analysis</i>	79
	<i>Association rules mining</i>	90
	<i>Analysing opinions through social media data mining (Twitter data analysis)</i>	92
	ANNEX II - ARGANDA USE CASE	95

Glossary and abbreviations

SOCS	Service-based Open Collaboration Space
CRFS	City Region Food Systems

AI	Artificial Intelligence
ICT	Information and Communication Technologies
KPI	Key Performance Indicators
UFSE	Urban Food Systems and Ecosystems
WP6	Work Package 6
D6.2	Deliverable 6.2
T6.2	Task 6.2
FAO	Food and Agriculture Organization
ML	Machine Learning
NLP	Natural Language Processing
API	Application Programming Interface
GIS	Geographic Information System
WMS	Web Map Service
WFS	Web Feature Service
OGC	Open Geospatial Consortium
FL	Federated Learning
GEMET	GEneral Multilingual Environmental Thesaurus
NER	Named Entity Recognition
CKAN	Comprehensive Knowledge Archive Network
ANN	Artificial Neural Network

1 Introduction

In this first Section, the context and global overview of Task 6.2 is described, with a special focus on document D6.2, which is the outcome from this task and collects the development and integration efforts of the

Deliverable D6.2

Prepared by P21 | Edited by P19, P20, P21, P35, P37 | Checked and reviewed by P01, P02 | Approved by P20
Rev 1.4 - November 2022

different datasets consumed and generated in the context of the Cities2030 project, into the proper subsystem of the S2CP platform developed in WP6.

1.1 Short description of the Cities 2030 project

The main goal of Cities2030 is to create a future proof and effective UFSE via a connected structure centered in the citizen, built on trust, with partners encompassing the entire UFSE. Cities2030 is committed to work towards the transformation and restructuring of the way systems produce, transport and supply, recycle and reuse food in the 21st century. Cities2030 vision is to connect short food supply chains, gathering cities and regions, consumers, strategic and complement industry partners, the civil society, promising start-ups and enterprises, innovators and visionary thinkers, leading universities and research across the vast diversity of disciplines addressing UFSE, including food science, social science and big data.

1.2 Short description of the WP6 Package

This work package will gather, design, and develop the main components and technological tools to establish a data-driven CRFS management platform for data collection, analysis and representation in multiple interfaces. An initial requirement acquisition will lead to the proposal of a common technical architecture for Cities2030, for with supporting data set will be incorporated to be considered for data analysis and representation. Particularly, a service-based open collaboration space will be incorporated, to be used by Cities20300 participants to improve their multistakeholder dialogue processes. In this space, blockchain technology will be employed to provide some proof of concepts of token-based monetization processes, and reflect multi-stakeholder interaction in a reliable and transparent way. Documentation and software repositories will be available for policy labs and living labs to develop their own solutions with assistance from WP6.

1.3 Purpose of D6.2 report

The objective of document D6.2 describes the work performed in Task 6.2, which **focuses on acquiring, preparing, analysing, fusing and integrating the incoming datasets that support the project use cases in a horizontal manner**, as well as provide the mechanisms to access and share them. In handling such large and heterogeneous datasets, **techniques such as distributed compression and compressed learning can be taken into consideration for effective data management**. Data mining methods such as clustering, and feature extraction will be investigated to prepare data conducting effective analysis. Distributed learning techniques such as **Federated learning and representative learning approaches can then be taken into consideration for developing collaborative data integration and analysis mechanisms**. Additionally, techniques of **prediction with guarantees such as causal and conformal learning will be investigated**.

1.4 Relation of the D6.2 Report with other WPs and deliverables

This deliverable collates the use cases (from D3.7), the requirements from the use cases (from D6.1), technical architecture (from D6.1), data-source definitions (from use-case analysis), data roles, security requirements (from T6.4), data governance considerations (from T6.4) and provides a design to be employed in Cities2030 for Data Integration and Management. The results of the data-analysis are available to the Data Visualisation component (T6.5) as described in D6.5.

Besides, this deliverable requires the specification of data models (from WP2, WP3, WP4 and WP5) in order to define data model profiles to be integrated into database. The framework design and implementation will be fully aligned with the work in WP8.

1.5 D6.2 structure

The structure of this deliverable is as follows:

Section 2 describes the work methodology considered for T6.2 (from WP6). Although this methodology refers only to the part of the S2CP platform architecture that has been developed in the context of T6.2, it is an approach that is fully compatible with the general working methodology of WP6 and with the CDM development methodology of the S2CP platform.

In Section 3, the state of the art is presented. Three big topics are addressed. First, a review of the state-of-the-art for Data Mining is provided. This covers the tasks, operations and techniques used theoretically to achieve results. It then reviews some of the available tools and technologies, with a view to identifying technologies that need to be interfaced with and technologies that can be reused. Second, it discusses the learning models and approaches used to derive trained models for use. It covers the challenges that need to be faced as it discusses some of the design decisions taken when providing a trained model for classification or inference. This second subsection also summarizes some of the key advantages of some of the approaches relative to one another. Finally, technologies developed within previous projects and that can be reused into the S2CP platform are analyzed. For example, Polirural project, DEMETER project and BusPrime project are considered.

In Section 4 the architecture and design for the data integration and analysis subsystem is presented. Basically, the subsystem includes five components and services: the data integration service, the geospatial information service, the sentiment and social network analysis component, the data mining component and the open data repository. This section includes two main subsections. First, the functionalities to be considered in the S2CP data subsystem are studied, considering the current literature and the state-of-the-art. On the other hand, the best approaches to implement those functionalities in the S2CP platform are analyzed, and finally the components that will be implemented to support the referred services are presented and described.

Section 5 presents the first of the components, aimed to analyzed social networks through sentiment analysis techniques. The objective is to implement techniques such as conformal learning or causal analysis to extract information from social networks describing how the user feel about their City-Region Food System. This component is specifically focused on identifying the users' reaction to some keywords provided by Living Labs.

Section 6 presents and describes the components focused on geospatial information. This component identifies the relevant point related to the CRFS in each one of the policy and living labs that are part of the Cities2030 project.

In Section 7 the data integration and analysis service are presented. In this section three main contributions are provided. First, a general description of the proposed services and functionalities is provided. Second, the general data model to be employed in the Cities2030 project is introduced. And, finally, the different data sources to be integrated into the S2CP platform are identified, described and analyzed.

Finally, in Section 8, the conclusions of the document and future lines of work are presented.

As additional content, Annex 1 gives a more detailed description of a use case from one of the Living Labs (Quart de Poblet). This analysis takes as an input the survey data provided by this lab and applies sentiment analysis techniques. Additionally, in Annex II, the Arganda use case is also described.

1.6 Roles and responsibilities

Lead partner role: WIT (P21) coordinates the task: provides guidance; steers implementation and secures alignment; and implements corresponding activities to deliver planned outcomes.

The rest of task participants: UPM (P20), Uni.lu (P35), ITC (P30) and SINNO(P19) develop the task simultaneously at EU level and beyond and contribute to the focus group.

2 Methodology for data integration and analysis

The objective of S2CP data integration and analysis subsystems the provision of technological solutions to improve the decision and analysis process of living and policy labs in the Cities2030 project. For this, a set of technological components capable of capturing, integrating and analysing relevant information is designed and implemented.

To carry out the implementation tasks, each one of the identified components will be led by one of the members of the task. A joint meeting will be held every two weeks, in which the status of the developments will be reviewed, those issues that must be transversal to the entire task or work package will be harmonized, and the task agenda for the next period will be proposed. The following table shows those responsible for each one of the components that are part of the data integration and analysis subsystem of the S2CP platform.

Table 1. Data integration and analysis components and responsible partners

Component	Geospatial information	Sentiment analysis	Open data repository	Data integration	Data mining
Responsible	UPM	UNIL	UPM	WIT	UPM

The components will be developed in parallel, with several development cycles in accordance with the CDM methodology and the schedule planned for WP6 (described in D6.1). In any case, in order to achieve the planned Development Objectives within the established times, a work methodology will be followed that we briefly describe in the following subsections.

2.1 First phase

In the first phase, a detailed study of the state of the art is carried out in relation to techniques and technologies for data integration and analysis. This is an ambitious work, given the large number of proposals that have been reported in the scientific literature and at a commercial level in the last 10 years. This detailed study, whose results can be consulted in Section 3, was structured in three different topics: data mining, learning and analysis techniques and data models (proposed in previous projects).

2.2 Second phase

In the second phase, the requirements captured in T6.1 will be considered to proceed with the design of the data integration and analysis subsystem, as well as the specific design of each of its components. The requirements and components identified within the general architecture (see D6.1) will be analysed among all the members of T6.2 in order to identify the software elements to be implemented, as well as their interrelationships. Section 4 presents the results of this design process

2.3 Third phase

In the third phase, development work will begin in accordance with the general CDM methodology. To this end, an agile development scheme will be followed, in which Labs will be interacted with on a regular basis, either through specific workshops for the general public or through individualized and personalized bilateral training meetings for each of the Labs.

In general, all these meetings focused on those components planned for development in the first macrocycle of the CDM methodology (see D6.1). These components are: Sentiment analysis (described in Section 5), Geospatial information (described in Section 6), and data integration and analysis (described in Section 7). These components present different levels of maturity, since they cover the first complete macrocycle (M1 – M18) and a first part of the second macrocycle (M18 – M36). In Section 8 we can find both a description of the real maturity level of the components, as well as a plan for future steps and developments.

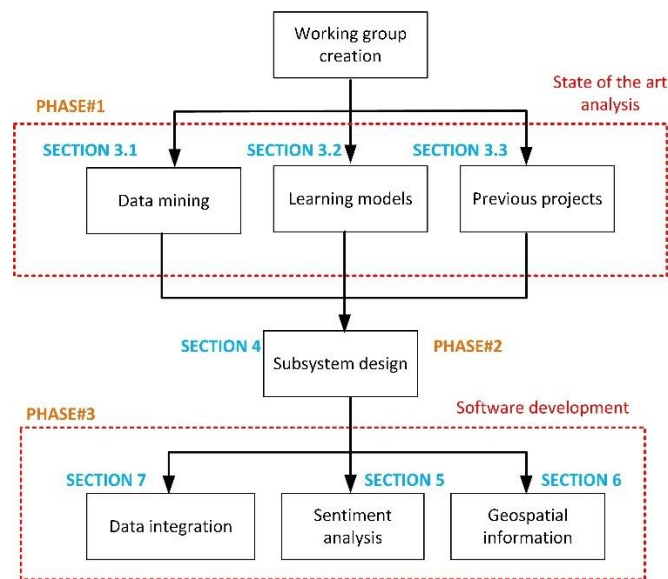


Figure 1. Methodology for the development of the data integration and analysis subsystem

3 State of the art

In this section we present the results extracted from the analysis of the state of the art on techniques and technologies for data integration and analysis.

3.1 Introduction

Data mining is the process of extracting and discovering patterns and other valuable information in large datasets. Broadly speaking, data mining is the computer-driven process of exploring data sets, identifying trends and anomalies, and subsequently analysing these findings to form conclusions and make better decisions. Data mining is applied in many industries to develop insights of the processes and study cases to improve efficiency and innovate on existing business models⁴.

⁴ Tan, P.-N.; Steinbach, M.; Karpapne, A.; Kumar, V. Introduction to Data Mining; Second edition.; Pearson: NY, 2019; ISBN 978-0-13-312890-1.

The data mining process applies methods from machine learning, statistics, and database systems to extract information from datasets into comprehensible structures that can be subsequently used to gather knowledge. The process involves operations such as data pre-processing, model and inference considerations, complexity considerations, post-processing of discovered structures, visualization, and updating. Given the evolution of data warehousing technologies and the growth of big data, adoption of data mining techniques has rapidly accelerated to transform raw data into useful knowledge, enabling organizational decision-making through insightful data analyses, although challenges related to scalability and automation are still faced. The data mining process is employed to generate descriptions and predictions about a target dataset and involves four main steps to extract valuable information from large data sets⁵:

1. **Setting the objectives:** The partners involved need to work together to define the studied task, to understand the context and to prepare the data questions and parameters.
2. **Data gathering and preparation:** Once the scope is defined, data scientists need to identify the data that will answer the posed questions correctly. Once the relevant raw data is collected, it will need cleaning (transformation, data migration, ETL (extract, transform, and load), removing outliers and noise, duplicates, dealing with missing values, data integration, and aggregation). Without this step, data is either meaningless or unreliable, due to its quality.
3. **Applying data mining algorithms** (building models and mining the patterns): Depending on the type of analysis, data scientists may begin with investigating data relationships, association rules, or correlations. In case of strong correlation between the features, a reduction of dimensionality is recommended to retain the most important features and achieve the best model performance. Deep learning algorithms are often applied to classify or cluster the data, depending on the available data. In this regard, if data is labelled, supervised learning techniques such as classification (to learn categories of data) or regression (to estimate the likelihood of an event) can be employed. If the data is not labelled, unsupervised learning techniques such as clustering can be applied to discover underlying, unknown similarities in characteristics considered.
4. **Evaluation of results and implementation of the knowledge:** Once the data is processed, the results need to be evaluated and interpreted by experts to provide valid, novel, useful, and understandable knowledge that can be considered in the implementation of new strategies and in achieving the defined objectives.

The data mining process (part of the Knowledge Discovery in Databases workflow, of KDD) is widely applied in organizations and industries to extract knowledge from available data⁶. Some of the most popular data mining use cases include operational optimization (to reduce costs across operational functions by identifying costly bottlenecks, enabling of a more efficient running of organizations and improving decision-making processes in education (to evaluate and optimize the performance the universities), sales and marketing (to optimize marketing campaigns and yield higher return on investment, or ROI), improving client segmentation, prospects and expectations using customer-related stored data), or detecting fraud (to assist financial institutions in detecting fraud, or eliminating fake user accounts).

⁵ Zaki, M.J.; Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms; Cambridge University Press: Cambridge, United Kingdom ; New York, NY, 2020; ISBN 978-1-108-47398-9.

⁶ Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings; Dong, Y., Kourtellis, N., Hammer, B., Lozano, J.A., Eds.; Machine Learning and Knowledge Discovery in Databases; Springer: Cham, 2021; ISBN 978-3-030-86516-0.

3.1 Data Mining Tasks, Operations and Techniques

Depending on the use case, data mining techniques are applied by data scientists throughout the data mining process to describe data to study patterns, associations, and correlations, to identify outliers for use cases, or to classify and cluster data through classification and regression methods. The data mining techniques can be divided into two main purposes; they can either describe the target dataset or they can predict outcomes using machine learning algorithms to turn large volumes of data into useful information⁷. Some of the most common operations are:

3.1.1 Data Warehousing

Data warehousing involves storing structured data in relational database management systems or semi-structured and unstructured data stores (NoSQL) in servers or in the cloud to enable its future in-depth analysis.

Traditionally, data warehouses were highly structured, and unified and contained locally stored information from a variety of data sources such as Customer Relationship Management (CRM), finance systems, Internet of Things, and many others. However, given the explosive data growth, the current trend is to dynamically grow or shrink data warehouses in the cloud to meet rapidly changing requirements.

3.1.2 Data Cleaning

Data cleaning (or pre-processing) is the process of preparing the available data for mining and involves operations such as organizing the data, removing duplicates, inaccurate or corrupted data, or dealing with the null values to minimize the impact that poor data quality has over the mining procedure. Some of the most common methods are:

- **Verifying the data and converting data types:** This task involves checking that the data is in the proper format and ensuring its uniformity.
- **Removing irrelevant data** to eliminate data points that are not necessary for the analysis.
- **Eliminating duplicate data points and errors** to boost efficiency and reduce the effect the incorrect data has over the outcomes of the analysis.
- **Completing missing values** to provide estimated values for missing data points. Please note that, if possible, it is highly recommended to eliminate registries containing missing values and increase the efforts to generate bigger, complete datasets.

3.1.3 Statistical Techniques

The statistical techniques represent the core of most techniques involved in the data mining process, as most analysis models are based on statistical concepts that output numerical values or describe the variables involved. Two of the most applied statistical techniques are:

- **Descriptive Statistics** to organize data, identify the main characteristics of the mined data, and summarize it into graphs or numbers. Some of the most common such commonly used descriptive statistical methods are computing the average, the mode, the standard deviation, or the correlation (to capture the relationship between variables in a pair).

⁷ Data Mining: Practical Machine Learning Tools and Techniques; Witten, I.H., Witten, I.H., Eds.; Fourth Edition.; Elsevier: Amsterdam, 2017; ISBN 978-0-12-804291-5.

- **Inferential Statistics** to draw conclusions based on probability theory and generalize the data. By analysing sample statistics, it is possible to infer parameters about populations and make models of relationships within data.

3.1.4 Data Visualization

Data visualisations represent an essential element of data mining and grant the users insights into data by revealing **different trends and patterns in data**, often in a dynamic, colourful way by means of dashboards. This operation involves presenting the data in charts, graphs, maps, or diagrams. The most popular methods for visually representing data are:

- **Comparison charts:** Charts and tables can be used to express relationships in the data (for example, monthly sales of a product over a one-year period).
- **Maps:** to visualize data from specific geographic locations (e.g., analyse characteristics and densities, detect trends, changes in performance between different regions).
- **Heat maps** are used to represent data through distinct colours and shading pallets to indicate patterns and ranges in the data.
- **Density plots** are used to represent occurrences of single events over time.
- **Histograms:** to organize a group of data points into user-specified ranges. Histograms condense data series by taking many data points and grouping them into logical ranges (bins).
- **Network diagrams:** to show how data points relate to each by connecting objects using lines (links).
- **Scatter plots** represent data point relationships on a two-variable axis to compare unique variables.
- **Word clouds** are used to highlight specific words or groups of words in a body of text (corpus) using the frequency.

3.1.5 Outlier Detection

The outlier detection operation is used to determine anomalies in datasets and understand why these anomalies happen and anticipate, optimize, control the behaviour, and prepare a response (e.g., peak traffic scenarios, fraud detection, maintaining safe databases, anomalies in their production). The operation is also employed to detect, find, and eliminate errors (incorrect data or wrong samples). The most applied methods are:

- **Numeric outlier:** Outliers are detected based on the Interquartile Range (the spread of the observations); datapoints outside that range are considered outliers.
- **Z-score** denotes how many standard deviations a data point is from the sample's mean (used for extreme value analysis).
- **DBSCAN**⁸: This stands for "density-based spatial clustering of applications with noise" and is a method that defines data as core points, border points, and noise points, which are the outliers.

3.1.6 Association

Association is applied to discover data points or events that are linked. The concept of association is related to the statistical concept of correlation and involves the analysis of data to show if relationships between data points exists. It works by finding all frequent items in the dataset and creating strong association rules

⁸ Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA; Simoudis, E., Han, J., Fayyad, U.M., Eds.; AAAI Press, 1996; pp. 226–231.

from the frequent items. Association rules help finding and understanding relationships between variables in a dataset and are particularly useful when developing recommendation engines or improving cross-selling. The most popular technique is:

- **Apriori Algorithm**⁹ finds the frequent sets of items in a dataset by an iterative process called candidate itemset generation. At the end, the algorithm generates candidate item sets and association rules using a measure called confidence (the occurrence of items together in a dataset).

3.1.7 Regression

Regression is applied to reveal the relationship between a dependent (target, continuous variable) and the **independent variable(s), or characteristic(s)**, and is widely used for forecasting, time series modelling, and finding the causal effect relationship between the variables. The most common regression techniques are:

- **Linear Regression:** establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). Multiple linear regression has $n > 1$ independent variables. If the power of an independent variable is more than one, the technique is called polynomial regression. The Least Square Method can easily accomplish these tasks.
- **Logistic Regression:** used to find the probability of event=Success and event=Failure and can be applied when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature, the value of the output ranging from 0 to 1.
- **Ridge Regression**¹⁰: is a regularization method and uses l2 regularization to solve the multicollinearity problem through a shrinkage parameter λ (lambda) and applies squared values in the penalty function.
- **Lasso Regression** (Least Absolute Shrinkage and Selection Operator): uses l1 regularization to reduce the by shrinking coefficients to zero (exactly zero) after applying absolute values in the penalty function, which helps in feature selection.
- **ElasticNet Regression:** a hybrid of Lasso and Ridge Regression techniques is trained with L1 and L2 prior as regularizations and encourages the group effect in case of highly correlated variables.

3.1.8 Classification

The task of classifications involves analysing the characteristics of the available data (stored in data warehouses) to categorize, or classify, related data. It has an important application in forecasting events using patterns in data (e.g., image recognition, fraud detection). The most popular methods for data classification are:

- **Logistic regression:** This algorithm attempts to show the probability of a specific outcome within two possible results (0 and 1, binary class). When applying the algorithm to classification tasks, a decision limit is drawn (usually 0.5), the class assigned to each data point being the closest (0 or 1) to the computed probability value.
- **Decision trees:** The data points are diagrammed into a chart called a decision tree to compute decision limits. Decision trees enable users to clearly understand how the input data affect the

⁹ Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile; Bocca, J.B., Jarke, M., Zaniolo, C., Eds.; Morgan Kaufmann, 1994; pp. 487–499.

¹⁰ Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 2000, 42, 80–86, doi:10.1080/00401706.2000.10485983.

output (a tree-like visualization to represent the potential outcomes of these decisions can be obtained). When various decision tree models are combined, they create a random forest (a basic form of ensemble modelling, often significantly more accurate than decision trees on their own, but more prone to overfitting). As a side note, this technique can also be applied to regression tasks (obtaining a value based on a set of decisions).

- **K-nearest neighbours (KNN)** assigns a label to data by comparing it to known similar points. It is a non-parametric algorithm that classifies data points based on their proximity and association to other available data. It works by calculating the distance between data points, usually through Euclidean distance (although other distances such as Manhattan, or Minkowski are used) to assign a category to the data, based on its characteristics.
- **Naive Bayes:** the algorithm is based on the Bayes Theorem of Probability¹¹ and uses tagged data to predict whether similar events will occur.
- **Support Vector Machine (SVM)**¹²: This machine learning algorithm is often used to define the line that best divides a data set into two classes and works by searching the decision line that maximises the differences between classes. In the first step, the algorithm represents the input instances as points in a multidimensional or hyperdimensional space (depending on the number of input features). Subsequently, the examples are mapped so that they are divided by a space as wide as possible. The points are then separated by a hyperplane that provides the maximum separation between classes (the decision limit, or Maximum Margin Hyperplane).

3.1.9 Clustering

Clustering techniques are applied to understanding data and display how the distribution of data is related to distinct types of metrics, mechanisms, or criteria. The data points are represented in clusters to identify relevant trends.

- **Partitioning method** involves dividing a data set into a group of specific clusters for evaluation based on a chosen criterion.
- **Hierarchical method:** Data points are grouped based on similarities to create clusters that can then be analysed separately.
- **Density-based method:** A method where data points are analysed using established processing mechanisms such as DBSCAN; data points that are isolated (being labelled "noise") and discarded.
- **Model-based method:** models are created for each data cluster to locate the data points that fit the model.

3.1.10 Predictive Modelling

The prediction is a powerful aspect of data mining where the patterns found in current or historical data are extended into the future, enabling insights into what trends will happen next in their data. Predictive modelling seeks to turn data into a projection of future action or behaviour. These models examine data sets to find patterns and trends, then calculate the probabilities of a future outcome. Predictive modelling is among the most common uses of data mining and works best with large data sets that represent a broad sample size. Predictive modelling uses techniques and terminology similar to other data mining processes. Although machine learning is widely applied in more advanced aspects of it, predictive analytics can also be

¹¹ https://en.wikipedia.org/wiki/Bayes%27_theorem

¹² Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* 1995, 20, 273–297, doi:10.1007/BF00994018

facilitated with more traditional statistical techniques. There are several different approaches to using predictive analytics, are four examples:

- **Time series modelling and Forecasting**¹³ is applied to analyse the data based on the temporal component of the input (e.g., study of Consumer Price Index trends over a year). Generally, the goal is to obtain a model capable of making predictions by modelling historical data (e.g., weather forecasting). This field is currently transitioning from the descriptive and parametric deterministic or stochastic (Box Jenkins Approach, ARIMA (Auto Regressive Integrated Moving Average)) towards the Machine Learning and Deep Learning approach. One powerful AI (artificial intelligence) technique commonly applied is long-term **memory processing** which enables the successful analysis of data from over extended periods of time and can identify patterns too subtle to detect otherwise.
- **Classification modelling** to place data into groups where it can be used to answer direct questions (some of the techniques were specified in Section 2.1.1.8).
- **Cluster modelling:** to obtain a predictive model can be used to study the clusters of groups with shared characteristics and make decisions (some of the techniques were specified in Section 2.1.1.9).

3.1.11 Machine Learning

Machine learning methods are part of the advanced developments in data mining that enables the computers to learn “on their own”. ML techniques are useful for processing semi-structured and unstructured data to quickly identify patterns and insights. ML methods have a wide range of applications, from predicting buying patterns, to optimizing campaigns, to improving diagnostic and treatment methods in healthcare. The most common techniques are:

- **Supervised learning:** models learn from input data labelled with correct values. It is called supervised because during the training process, the algorithm has to iteratively process labelled data to predict the outcome and improve its performance. Supervised machine learning is extensively used in classification and regression operations.
- **Unsupervised learning:** models process unlabelled data to identify underlying hidden patterns. This type of learning is widely applied to perform clustering and association.
- **Semi-supervised learning:** a hybrid machine learning technique where models are trained using a combination of labelled and unlabelled input data.
- **Reinforcement learning:** The models learn to make decisions in game-like situations, generally by maximizing a reward function. In this approach, the data is examined in specific complex environments, the actions applied by the algorithm following the “trial and error” concept until the models find a solution to the problem.

3.1.12 Neural Networks and Deep Learning

Deep learning algorithms model the input, training data by means of layers of nodes (also called “neurons”). Each neuron is made up of inputs, weights, a bias, and an output. If the output value exceeds a threshold activation value, the data is passed to the next layer in the network. The most basic type of neural network is called perceptron and consists of three main layers: input, “hidden,” and output. Here, the data enters the network through the input layer, is processed in the hidden layer, and is resolved in the output layer, where any relevant action based on the data is taken. The hidden layer can consist of many processing layers (multilayer perceptron), depending on the amount of data being used and learning taking place. This field is

¹³ PEIXEIRO, M. Time Series Forecasting in Python; O'REILLY MEDIA: S.I., 2022; ISBN 978-1-61729-988-9.

rapidly evolving towards the proposal of very deep neural networks and novel learning structures that enable a more efficient and higher quality learning procedure.

Neural networks are extensively trained with supervised learning techniques, have a high generalization capacity and are capable of learning complex mapping functions between the input and output by minimising the loss function through stochastic gradient descent procedures to deliver high performance metrics on novel, unseen data. Although a neural network can be a powerful tool in data mining, some of these neural network models are very complex, which makes it difficult for understanding how an output was determined ("black-box" approach).

Deep learning enables the approach of more complex tasks such as computer vision, speech recognition, or text analytics (using Natural Language Processing) and offers highly accurate predictions when large amounts of data are available. However, each of these complex tasks features specific types of neural network architectures or learning workflows for achieving the desired models.

3.1.13 Text mining

Text mining is a process of extracting useful information and patterns from a large volume of text (called corpus) to mine the text and find important data for the prediction and decision-making process. The text mining operations feature specific information retrieval techniques to deal with unstructured, heterogeneous, and more diverse nature of the input text data¹⁴.

In this regard, standard text mining methods discover lexical and syntactic features in the text and searches for patterns and relationships of data to turn unstructured text documents (text documents, emails, social media posts, multilingual texts, abbreviations, or slang) into valuable structured information¹⁵.

Some of the most important text mining applications are information extraction, sentiment analysis, semantic search, and conversational agents (bots). Text mining workflows are applied in research, life science (text regarding patient records), social media (monitor and investigate online content), business intelligence (analyse the customers and competitors to make better decisions), telecom sector, or commerce.

The main operations and techniques applied in text mining are:

- **Summarization:** To automatically extract partial content reflection (maintaining the main idea) from the whole content.
- **Information Extraction and Retrieval:** the process of extracting meaningful words from documents and relevant and associated patterns according to a given set of words or text documents (pattern matching formats).
- **Visualization** to represent information and visualize relationships using computer graphics.
- **Categorization:** To assign a category to the text among categories predefined by users. Supervised learning technique that categorizes the document according to content.
- **Clustering:** To segment texts into several clusters, depending on the relevance (grouping text according to similar characteristics, an unsupervised technique).
- **Natural Language Processing**¹⁶: tackling the automatic processing and analysis of unstructured text information (extracting meaningful information from text). The goal is to help define accurate and complete domain-specific taxonomies, always considered the high complexity of the natural language.

¹⁴ Handbook of Natural Language Processing; Indurkha, N., Damerau, F.J., Eds.; Chapman & Hall/CRC machine learning & pattern recognition series; Chapman & Hall/CRC: Boca Raton, FL, 2010; ISBN 978-1-4200-8592-1.

¹⁵ Sarkar, D. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing; Second edition.; Apress: New York, NY, 2019; ISBN 978-1-4842-4353-4.

¹⁶ Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python; 1st ed.; O'Reilly: Beijing ; Cambridge [Mass.], 2009; ISBN 978-0-596-51649-9.

- **Sentiment Analysis** (also called opinion mining) to identify a person's mood regarding a service or product. This operation has a high applicability in monitoring online reputation and increasing loyalty with customers
- **Question Answering** Using a list of patterns to answer a natural language question (e.g., developments of bots).

3.1.14 Web Mining

Web mining is a process on pattern identification from web data and involves the use various data mining techniques to extract knowledge related to the web content, structure, and usage (conducted on data from the web)¹⁷. It can be used for discovering useful information previously unknown and some of its most important applications are web advertising, e-commerce personalization, web-spam filtering, or web page categorization. Web mining can be classified into the following categories:

- **Web Content Mining:** the process of extracting data from web pages to search different patterns trends that gives useful insights. There are various techniques to extract useful data like web scraping which involves gathering or extracting useful location-specific information from the web (e.g., about a location, hashtag, keyword, event). This process is also called "web scraping" and is applied to collect and generate large data sets (e.g., tweet scraping, website scraping, scraping using official APIs).
- **Web Structure Mining:** search for new patterns and trends by extracting and analysing data from an individual source through multiple hyperlinks linked to the nested pages (that lead to other pages).
- **Web Usage Mining:** the process of extracting useful patterns, trends and information related to web page usage to discover designs and insights in web information (e.g., web-based documents and services, server logs, and hyperlinks).

3.2 Data Mining Tools and Technologies

The wide range of techniques available for data mining tasks needs appropriate tools to optimize the applied analytics processes. The techniques described in Section 3.1.2 usually require several tools for a proper execution, and for this reason, it is essential that the experts involved in the data mining process are able to identify the appropriate technologies. Additionally, in many cases, it is required to work with cloud data stores to perform analytics and obtain dashboards and data visualizations that provide the users with the information required to understand the knowledge. Next, a description of some of these tools and technologies, grouped by their application, will be provided.

3.2.1 Data Warehousing Technologies

TAs mentioned in Section 3.1.2, data warehouses are used to store and manage large amounts of historical data that would enable and support analytics. The necessary data for analytics is usually obtained by means of queries from the data warehouses and data lakes.

¹⁷ Mitchell, R.E. Web Scraping with Python: Collecting More Data from the Modern Web; Second edition.; O'Reilly Media: Sebastopol, CA, 2018; ISBN 978-1-4919-8557-1.

3.2.1.1 PostgreSQL

PostgreSQL¹⁸ is a free and open-source relational database management system (RDBMS) focused on offering scalability for complex data workloads and complying with SQL (Structured Query Language). It is used as the primary data store for many web, geospatial, and analytics applications.

3.2.1.2 MariaDB

MariaDB¹⁹ is also an open-source RDBMS that maintains high compatibility with MySQL.

3.2.1.3 MongoDB

MongoDB²⁰ is an open-source NoSQL database management program (NoSQL is used as an alternative to traditional relational databases that is useful for working with large sets of distributed data due to its flexible data scheme). MongoDB is a cross-platform document-oriented database tool that can manage document-oriented information, storing or retrieving the information using JSON-like documents. MongoDB provides high availability with replica sets, scales horizontally using sharding, and enables the definition of complex aggregation pipelines.

3.2.1.4 Cassandra

Cassandra²¹ is an open-source NoSQL database management system maintained by Apache Software Foundation. It is designed to manage large amounts of data across multiple distributed servers and provides high availability with no single point of failure. Unlike MongoDB, Cassandra is a column family database that stores a row with all its column families together—different column families are stored separately in a node, or in different nodes. Columnar databases are often used for reporting. Data models based on columns enable aggregate queries and store and process data by columns, instead of rows (resulting in significantly faster running times).

3.2.1.5 Apache Universe

The tools and technologies developed by Apache Software Foundation are distributed as open-source software to develop data analytics extracting relevant insights. The projects are particularly important in the world of data mining, due to their role in designing systems and data pipelines. Three of the most important technologies provided by Apache are described next.

¹⁸ PostgreSQL Documentation Available online: <https://www.postgresql.org/docs/9.1/biblio.html> (accessed on 28 February 2022).

¹⁹ MariaDB - Open Source Database (RDBMS) Available online: <https://mariadb.com/docs/reference/> (accessed on 28 February 2022).

²⁰ MongoDB Manual — Database References Available online: <https://docs.mongodb.com/manual/reference/database-references/> (accessed on 28 February 2022).

²¹ Shah, J.K.; Marin-Rimoldi, E.; Mullen, R.G.; Keene, B.P.; Khan, S.; Paluch, A.S.; Rai, N.; Romanielo, L.L.; Rosch, B.; Thomas W. Yoo; Maginn, E.J. Cassandra: An Open Source Monte Carlo Package for Molecular Simulation. *Journal of Computational Chemistry* 2017, 38, 1727–1739.

3.2.1.5.1 Hadoop

Hadoop²² is a collection of open-source libraries (library ecosystem) built to process large data sets. Each library performs its own specific tasks, making Big Data storage and processing more efficient. The framework is very versatile and allows most specialized Big Data software to connect to it. Hadoop has removed the need of relying on expensive hardware for high availability, achieving it through the software itself (which also detects points of failure). Hadoop's two main components address the problems posed by storing and processing Big Data:

- **Hadoop Distributed File System (HDFS)**²³: HDFS is Hadoop's distributed file system with automatic data redundancy and features a master/slave architecture. Hadoop is designed to work with very large files (GB, TB, ...) that are stored in very large block sizes (64 MB by default) and provides a batch-oriented data access.
- **MapReduce**²⁴— is a simple and versatile programming paradigm for parallel processing within the Hadoop framework and a core component, essential for Hadoop's operation. MapReduce is used to access large data stored in the HDFS file system and makes concurrent processing easy by breaking petabytes of data into smaller chunks and processing them in parallel on core Hadoop servers.

The support of the community and the quality of the implemented technologies have represented an important motivation in the adoption of Hadoop by companies (there are tools for data importing, monitoring, administration and coordination systems, or visual management interfaces that allow data analysts to decide when and how to perform these transformations, depending on the type of business questions).

3.2.1.5.2 Spark

Apache Spark²⁵ emerges due to the inefficiency of existing tools to handle data when executing iterative algorithms and data mining processes. Similar to Hadoop, Spark also focuses on processing data in parallel across a cluster, but runs in the random-access memory (RAM), resulting in significantly faster interactive queries (by keeping the data in memory, the overall performance is greatly improved). Spark is capable of processing data from all types of sources (local files, HDFS, Cassandra, MongoDB, Hive, PostgreSQL, MySQL, S3 Amazon, etc.), and compared to Hadoop, it is much faster, as data does not need to be retrieved in each iteration, being saved in memory (up to forty times faster than Hadoop).

Spark is based on the Resilient Distributed Dataset (RDD)²⁶ —a collection of elements partitioned across cluster nodes that can be operated in parallel. RDDs are created by starting with a Hadoop File System (HDFS) file and transforming it—a distributed collection only contains data of the same type (e.g., RDDs of strings, RDDs of integers, etc.). Once the data has been read as RDD objects in Spark, various operations can be performed through its APIs (Spark is developed in Scala, but there are three other APIs for programming: Python (PySpark), Java (Spark Java), and R (SparkR)). RDDs support two types of operations:

²² Apache Hadoop, Available online: <https://hadoop.apache.org/> (accessed on 28 February 2022).

²³ HDFS Architecture Guide Available online: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (accessed on 28 February 2022).

²⁴ MapReduce Tutorial Available online: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (accessed on 28 February 2022).

²⁵ Apache Spark - Unified Engine for Large-Scale Data Analytics Available online: <https://spark.apache.org/> (accessed on 28 February 2022).

²⁶ Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauly, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. In Proceedings of the Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012, San Jose, CA, USA, April 25-27, 2012; Gribble, S.D., Katabi, D., Eds.; USENIX Association, 2012; pp. 15–28.

- **Transformations** – to create a new dataset from an existing one (after applying a transformation, a new, modified RDD based on the original is obtained). For example, the “Map” operation passes each element of the dataset through a function and returns a new RDD representing the results of the transformation.
- **Actions** – return a value after executing a calculation on the data set (an action simply consists of applying an operation on an RDD and obtaining a value as a result, which will depend on the type of operation). For example, the “Reduce” operation can add all the elements of the RDD using some function and return the final result of the action.

3.2.1.5.3 Apache Hive

Apache Hive²⁷ is a data warehouse infrastructure built on top of Hadoop to provide Big Data aggregation, query, and analysis using SQL-like queries. Hive was created to allow users familiar with SQL to work with Big Data, using an interface called HiveQL. It is intended for the storage of very large databases (hundreds of TB) and enables users to read, write, and manage petabytes of data using SQL and perform queries on large volumes of data. Unlike relational databases (designed for interactive queries on small to medium data sets), Hive uses batch processing and is fast even on very large, distributed databases.

3.2.1.5.4 Apache Spatial Information System (SIS)

Apache Spatial Information System (SIS)²⁸ for geospatial data is a toolkit that handles spatial data to create interoperable geospatial applications and geospatial data representations. It provides a unified metadata model for file formats used for real-time smart city visualization, geospatial dataset discovery, or location-enabled emergency management.

3.2.1.5.5 Other Apache Tools and Technologies

- **Kafka**²⁹: is a distributing stream platform that allows publishing and subscription of record streams and enables real-time data pipelines (to react and transform according to a real-time ingested data stream).
- **Zookeeper**³⁰: is a centralized service used by distributed systems to maintain configuration information. It acts as a key-value store that is hierarchical for storage or managing of data, its main purpose being the synchronization of applications.
- **Lucene**³¹: is a library of search engine software that supports the search and facilitates the indexing platform. It is commonly used for the processing of ranked searching.

²⁷ Camacho-Rodríguez, J.; Chauhan, A.; Gates, A.; Koifman, E.; O'Malley, O.; Garg, V.; Haindrich, Z.; Shelukhin, S.; Jayachandran, P.; Seth, S.; et al. Apache Hive: From MapReduce to Enterprise-Grade Big Data Warehousing. In Proceedings of the Proceedings of the 2019 International Conference on Management of Data; ACM: Amsterdam Netherlands, June 25 2019; pp. 1773–1786.

²⁸ Apache SIS - The Apache SISTM Library Available online: <https://sis.apache.org/> (accessed on 28 February 2022).

²⁹ Garg, N. Apache Kafka; 2013; ISBN 978-1-78216-793-8.

³⁰ Hunt, P.; Konar, M.; Junqueira, F.P.; Reed, B. ZooKeeper: Wait-Free Coordination for Internet-Scale Systems. In Proceedings of the 2010 USENIX Annual Technical Conference, Boston, MA, USA, June 23-25, 2010; Barham, P., Roscoe, T., Eds.; USENIX Association, 2010.

³¹ Apache Lucene Available online: <https://lucene.apache.org/index.html> (accessed on 28 February 2022).

3.2.2 Python Scientific Stack

In most data mining workflows, the data preparation operation represents an important part of projects, the efficient storage, loading, and correct manipulation of numerical arrays being fundamental in data analysis processes. The Python Scientific Stack contains a series of open-source libraries focused on the manipulation of vectors and matrices, where the common operations are already encapsulated as functions and vectorized. At its core, it includes a general-purpose programming language (Python^{32,33}), an interactive computing environment (e.g., Jupyter) and a set of libraries for scientific computing (processing, visualization, statistics, etc.), some of which will be described next.

3.2.2.1 Anaconda

Anaconda³⁴ is a distribution that includes Jupyter³⁵ (a server kernel; an interactive computing protocol). Anaconda comes with its own package manager, called **conda**, which can be used as a replacement or as a complement to **pip** to install libraries and avoid conflicts in the requirements (as it has the ability to create isolated environments).

3.2.2.2 NumPy

NumPy³⁶ is an open-source library specialized in working with arrays (matrices) that can come from other data and have different dimensions. The data must be homogeneous (data of a single type, e.g., 8-bit or 32-bit integers, but not a mix of the two), and once the array is built, a large number of pre-implemented functions can be applied to its elements. NumPy's mathematical operations are optimized by vectorizing arrays (compiled in C resulting in processing up to one thousand times faster than Python) and feature functions for working in the field of linear algebra, Fourier transforms, and matrices. It also contains functions to concatenate, slice, split, sort, reshape arrays, etc.

3.2.2.3 Pandas

Pandas³⁷ is an open-source data manipulation and analysis tool widely used for data wrangling operations. "Data Wrangling" is the process of transforming and processing input data to make it more suitable for a variety of downstream purposes, such as analysis. This can include data visualization operations, data aggregation, training a statistical model, as well as many other potential uses.

³² Python Documentation Available online: <https://docs.python.org/3/> (accessed on 28 February 2022).

³³ Van Rossum, G.; Drake, F.L. Python 3 Reference Manual; CreateSpace: Scotts Valley, CA, 2009; ISBN 1-4414-1269-7.

³⁴ Anaconda Inc. Anaconda Software Distribution Available online: <https://docs.anaconda.com/>.

³⁵ Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks – a Publishing Format for Reproducible Computational Workflows. In Proceedings of the Positioning and Power in Academic Publishing: Players, Agents and Agendas; Loizides, F., Schmidt, B., Eds.; IOS Press, 2016; pp. 87–90.

³⁶ Van Der Walt, S.; Colbert, S.C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 2011, 13, 22.

³⁷ McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the Proceedings of the 9th Python in Science Conference; Walt, S. van der, Millman, J., Eds.; 2010; pp. 56–61.

The “pandas” library provides some powerful tool sets like “DataFrame” and “Series”, based on with tabular concepts that can store, process, and analyse heterogeneous data (coming from different data formats), while NumPy offers an object called Array for number crunching which only works with numerical data.

3.2.2.4 Matplotlib

Matplotlib³⁸ is an open-source library for creating plots that allows the control of all the details of static graphics. Matplotlib allows the saving of plots in different vector and raster image formats, including PDF, JPG, SVG, PNG, BMP, and GIF. The development of the library is focused on stability and evolves slowly to incorporate elements from other libraries—there are many other plotting libraries that are built upon matplotlib.

3.2.2.5 Statsmodels

The purpose of data processing is to test hypotheses and there are open libraries for studying distributions, statistical inference, and linear models (e.g., SciPy). Statsmodels³⁹ is a Python library that allows to estimate statistical models and perform statistical tests; the part of the Python scientific stack that is geared towards data analysis, data science, and statistics. It is built on top of the NumPy and SciPy numerical libraries and integrates with Pandas for data handling and contains an extensive list of descriptive statistics, test statistics, plotting functions, and output statistics for different types of data.

3.2.2.6 Seaborn

Seaborn⁴⁰ is built on top of matplotlib, integrates with pandas, and uses statsmodels library for the statistical part. The basic idea behind is the grouping of visualizations with certain types of common statistical analysis and includes additional utilities such as “style palettes”. Seaborn also integrates linear models.

3.2.3 Dashboard Creation – Business Intelligence (BI) Technologies

A dashboard is a page, often called a canvas, where a “story” is told through visualizations. These visualizations are derived from reports, and each report can be based on a distinct set of data. A dashboard is limited to one page, should be well designed, contain only the highlights of that “story”, and allows readers to view specific reports to get more details. A dashboard could be considered the “gateway” to the underlying reports and datasets (selecting a visualization will take us to the report, and the dataset on which it is based). BI tools are simple enough for most business users to use without requiring special programming skills and facilitate the measurement of organizational performance, while increasing the value of organizational information and data.

3.2.3.1 Power BI

Power BI⁴¹ is a commercial business intelligence and analytics tool (and service) used to manage, analyse, visualize, and share data. It is capable of unifying data from many sources and creating interactive dashboards

³⁸ Hunter, J.D. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 2007, 9, 90–95, doi:10.1109/MCSE.2007.55.

³⁹ Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference; 2010.

⁴⁰ Waskom, M. Seaborn: Statistical Data Visualization. JOSS 2021, 6, 3021, doi:10.21105/joss.03021.

⁴¹ Data Visualization | Microsoft Power BI Available online: <https://powerbi.microsoft.com/en-us/> (accessed on 28 February 2022).

and reports. Gartner confirmed Microsoft as a leader in "Analytics and Business Intelligence Platforms 2019," as a result of Power BI's capabilities. It is developed by Microsoft (not open source and only available for Windows 10 and Windows 11) and features a very intuitive interface that is suitable for non-technical users. Power BI uses DAX as the expression language, which includes a collection of functions, operators, and constants that can be used in a formula to define calculations and return one or more values in a similar syntax to Excel expressions.

3.2.3.2 Tableau

Tableau⁴² is a software suite that develops interactive data visualization products that focus on business intelligence. Tableau offers an intuitive interface for generating diverse types of charts using large, multidimensional relational databases, cloud databases, and traditional spreadsheets and contains mapping functionalities, where any latitude and longitude coordinates can be plotted (also offers custom geocoding).

3.2.3.3 Pentaho

Pentaho is free business intelligence software that provides data integration, multidimensional data services, reporting, information dashboards, data mining and extract, transform, load (ETL) capabilities. Its most important applications are:

- **Pentaho Data Integration**⁴³ provides Extract, Transform, and Load (ETL) capabilities that facilitates the process of capturing, cleansing, and storing data, being commonly used for data migration, loading big datasets, data cleaning, or data warehousing.
- **Pentaho Dashboard Designer**⁴⁴ enables the creation of panels (dashboards) intuitively and the analysis of several reports simultaneously. It is capable of generating dynamic charts and graphs within one space while creating reports in another. The application features dynamic filter controls that allow dashboard viewers to change the details of a dashboard by choosing different values from a list and control the content in one pane of a report by changing options in another.

3.2.4 Machine Learning (ML) and Deep Learning (DL) Frameworks

Machine learning frameworks provide interfaces that allow specialists to build, train and deploy models and are being used in almost all industries. These frameworks feature essential operations for tasks specific to model training such as data pre-processing, gradient descending, parameter optimization. The frameworks can deliver data-based models with high generalization capacities that achieve high performance metrics (depending on the task they were trained for). The three ML and DL frameworks described next have demonstrated their efficiency in scaling, training, and deployment ML lifecycles.

⁴² Tableau - Business Intelligence and Analytics Software Available online: <https://www.tableau.com/> (accessed on 28 February 2022).

⁴³ Pentaho Data Integration - Hitachi Vantara Lumada and Pentaho Documentation Available online: https://help.hitachivantara.com/Documentation/Pentaho/7.1/0D0/Pentaho_Data_Integration (accessed on 28 February 2022).

⁴⁴ Pentaho Dashboard Designer - Hitachi Vantara Lumada and Pentaho Documentation Available online: https://help.hitachivantara.com/Documentation/Pentaho/8.0/Products/Dashboard_Designer (accessed on 28 February 2022).

3.2.4.1 Scikit-learn

Scikit-learn⁴⁵ is an open-source library for the Python programming language that implements different machine learning methods. The library features extensive pre-processing capabilities, various classification, regression, and clustering algorithms for machine learning tasks (described in Sections 2.1.1.5-2.1.1.9) and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

3.2.4.2 TensorFlow and Keras

TensorFlow⁴⁶ is an open-source framework designed to solve machine learning problems with a strong focus on deep neural network training and inference. It provides functionalities that enable an easy implementation of the operations required to perform complex parallel computations and to build advanced neural network models. With its version 2.0, introduced in 2019, TensorFlow incorporated the popular Keras⁴⁷ library (modular and extensible, intended for beginner users) as its official high-level API (Application Programming Interface).

3.2.4.3 PyTorch

PyTorch⁴⁸ is an open-source deep learning-oriented framework created to offer similar production optimizations to TensorFlow but eases the writing of dynamic models.

3.2.5 Tools and Technologies Specific to Certain Data Mining Operation

In addition to the widely known tools and technologies described above, developers are also providing open-source software and solutions for particular use cases.

3.2.5.1 Gephi

Gephi⁴⁹ is open-source software specialized in visualizing and analysing large networks, widely used for exploring, analysing, filtering, clustering, manipulating, and exporting graph data. It can be used in combination with other technologies to analyse specific events (e.g., hashtag-scraped Twitter data) and examine certain behaviours within the digital world.

⁴⁵ Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830.

⁴⁶ Martín Abadi; Ashish Agarwal; Paul Barham; Eugene Brevdo; Zhifeng Chen; Craig Citro; Greg S. Corrado; Andy Davis; Jeffrey Dean; Matthieu Devin; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015.

⁴⁷ Chollet, F. Keras Available online: <https://github.com/fchollet/keras> (accessed on 14 May 2020).

⁴⁸ Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp. 8024–8035.

⁴⁹ Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open-Source Software for Exploring and Manipulating Networks. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009*, San Jose, California, USA, May 17-20, 2009; Adar, E., Hurst, M., Finin, T., Glance, N.S., Nicolov, N., Tseng, B.L., Eds.; The AAAI Press, 2009.

3.2.5.2 NLTK

NLTK⁵⁰ is an open-source library specialised in text mining processes that automates operations such as tokenization, phrase segmentation, lexical or morphological analysis, syntactical analysis, or semantic analysis in document collections. It is used in knowledge discovery in text processes to process, enhance, and enrich available text data.

3.2.5.3 Semex tool (PoliRural project)

There are also a wide variety of open tools and technologies resulting from publicly funded projects. One such example is Semex⁵¹, a web app based on text mining capable of extracting information from unstructured data and displaying the results through graphical and textual representations to support European rural areas in responding to contemporary challenges.

The tool implements natural language processing operations such as topic extraction, named entity recognition and sentiment analysis and combines them with geographical representations to enable graphical insights of rural-specific topics (12 pilots were implemented in different regions of Europe and Israel). These high-level apps are focused on providing an accessible use to a wide range of users (user-friendly).

3.2.6 Cloud-Oriented Solutions

In the last decade, cloud-oriented services for storage and processing are having a significant impact on the growth of data mining. We believe that, as the growth rate of data is increasing exponentially, the cloud is becoming the most appropriate place to store and process data. Consequently, we believe that data mining approaches will rely even more on the cloud and that future developments in cloud computing will be focused on incorporating more effective data mining tools.

In these cloud services, data is transmitted, stored, and processed on remote storage systems in exchange of a per-consumption rate (with capacity increasing and decreasing as needed). These technologies are suited for fast processing of Big Data (large amounts of structured, semi-structured and unstructured data), provide an easy scalability of the resources to meet the demands, and feature advanced forms of data mining like ML and DL in the cloud. The most important cloud infrastructures are maintained and managed by big tech companies such as Amazon, Microsoft, or Google.

3.2.6.1 Amazon Web Services

Amazon Web Services (AWS) provide on-demand cloud computing platforms and APIs and include a variety of distributed computing infrastructures and tools. Some of its most important services for data mining are:

- **Amazon S3**⁵² is a storage service focused on scalability, data availability, security, and performance that allows the storing of big data lakes.
- **Amazon Redshift**⁵³ is a cloud data warehouse service capable of efficiently storing and analysing petabytes-scales of structured and semi-structured data. The SQL dialect is based on PostgreSQL (described in Section 2.1.2.1.1).

⁵⁰ Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python; 1st ed.; O'Reilly: Beijing ; Cambridge [Mass.], 2009; ISBN 978-0-596-51649-9.

⁵¹ SEMEX - POLIRURAL Hub Available online: <https://hub.polirural.eu/semex> (accessed on 28 February 2022).

⁵² Amazon S3 API Reference - Amazon Simple Storage Service Available online: https://docs.aws.amazon.com/AmazonS3/latest/API/Type_API_Reference.html (accessed on 28 February 2022).

⁵³ Gupta, A.; Agarwal, D.; Tan, D.; Kulesza, J.; Pathak, R.; Stefani, S.; Srinivasan, V. Amazon Redshift and the Case for Simpler Data Warehouses. In Proceedings of the Proceedings of the 2015 ACM SIGMOD International Conference on

- **Amazon Machine Image (AMI)**⁵⁴: delivers pre-configured environments to quickly build deep learning applications and provides the experts with the infrastructure and tools to accelerate deep learning in the cloud, at a large scale. For example, Amazon EC2 instances can be easily launched with pre-installed popular deep learning frameworks to train complex, custom AI models, experiment with new algorithms, or learn new skills and techniques.

3.2.6.2 Microsoft Azure

Microsoft Azure is a cloud computing service operated by Microsoft via data centres that, among others, provides storage and data management services. One of its most important applications for data mining is:

- **Microsoft Azure Synapse**⁵⁵: a service that combines data warehousing and big data analytics to ingest and prepare data for business intelligence and machine learning needs. It features a distributed SQL processing engine for data warehousing workloads and is capable of horizontal and vertical scaling of the workflows.

3.2.6.3 Google Cloud

Google Cloud offers cloud services for data storage and computing. One of its most important applications for data mining is:

- **Google BigQuery**⁵⁶: a data warehouse that automatically scales and enables spatial analysis processes using SQL and Geographical Information Systems (GIS). In addition, it can build and train ML models on large-scale structured or semi-structured data and supports real-time interactive dashboarding.

3.3 Learning Models and approaches

ML is finding ways into our everyday life as more service providers attempt to use ML to gain insights into the data they already have. ML however relies on the tremendous amount of data generated by mobile and IoT devices. The more "traditional" approach for how to handle all this data is to stream the data to the cloud. In these cloud services, data is analysed, features are extracted and the corresponding machine learning models are trained.

Independent of the type of data mining or machine learning technique applied, some challenges permeate implementations, due to the centralised approach implied under classical machine learning.

3.3.1 Challenges

Under a centralised approach, data is collected and aggregated for training purposes in a small number of centralised nodes. This concentration of data in a small number of nodes has associated challenges:

- **Connectivity**: For example, in a large IoT system, data must be transmitted over a stable connection to the aggregating nodes. This implies power consumption to maintain a stable connection.

Management of Data; ACM: Melbourne Victoria Australia, May 27 2015; pp. 1917–1923.

⁵⁴ Amazon Deep Learning AMIs Available online: <https://aws.amazon.com/machine-learning/amis/> (accessed on 28 February 2022).

⁵⁵ WilliamDAssafMSFT, Azure Synapse Analytics - Azure Synapse Analytics Available online: <https://docs.microsoft.com/en-us/azure/synapse-analytics/> (accessed on 28 February 2022).

⁵⁶ BigQuery: Cloud Data Warehouse, Available online: <https://cloud.google.com/bigquery> (accessed on 28 February 2022).

- **Bandwidth:** The transmission of data consumes available connection bandwidth. Therefore, as the number of sources for data expands the amount of data generated also expands. For example, an electrical substation could generate 5 GB/s of data.
- **Latency:** As networks are not perfect there are delays in sending information to the centralised nodes. This makes supporting near real-time applications, where a decision has to be made within a small time frame a challenge.
- **Privacy:** Inherent in the centralised approach is the transfer of data across the network to the centralised nodes. This increases the attack surface, as the data must traverse the intermediate nodes. Thus this is problematic for sensitive operational data, where ideally such data should remain on site.
- **Transparency / fairness:** A small number of organisations (usually just one) that are aggregating the data, implies transparency problems. For example, how could another organisation verify that there is no bias (against any subgroup) in the algorithm used, without access to the data. It becomes very difficult to independently verify the trained algorithm is doing what it's supposed to do, or even verify that the algorithm that is in use matches the algorithm claimed to be in use by the data aggregator.

3.3.2 Distributed Learning

Edge/fog computing has addressed the limitations of cloud computing, especially the provisioning of fast response time to applications with strict latency requirements. It did so applying a Distributed Learning approach. Rather than sending both the data and a request to a centralised service, a pre-trained model is distributed instead. This pre-trained model is further refined by training it further using local data. This is shown in the following figure.

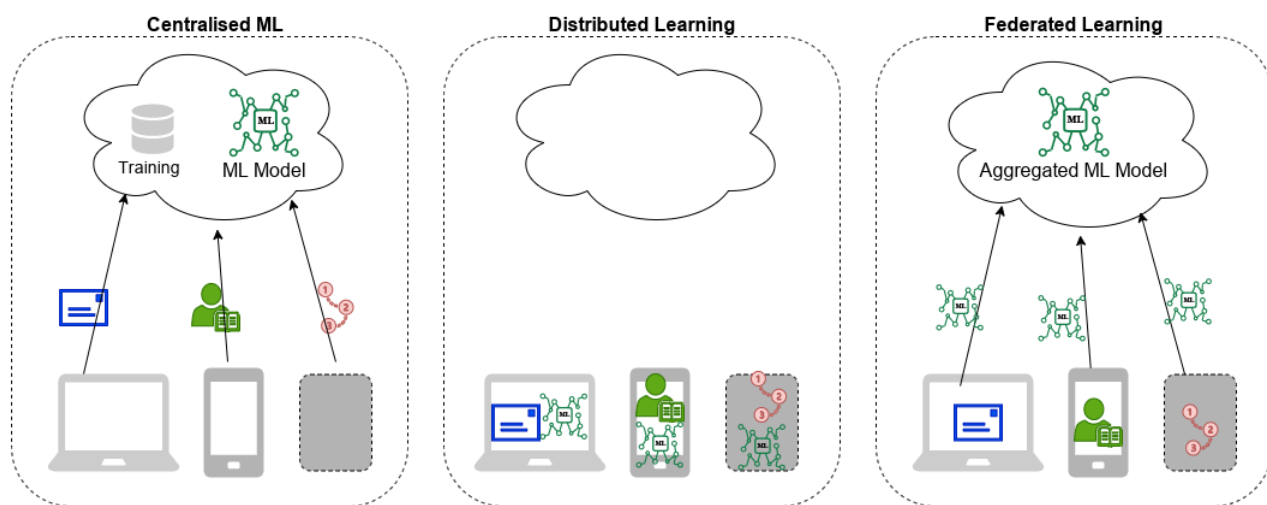


Figure 2. Centralised versus Distributed versus Federated Learning

Each device maintains its own model and trains it based on the locally available data. The trained model is not exchanged with peer devices. This approach improves the privacy outlook as data can stay on the device. However, because data is not shared then the performance of the ML model is degraded as it only has access to its own data for training purposes. It is not possible for it to learn from the data of its peers.

There have been a couple of application areas where this has been successfully applied, for example, smart classrooms⁵⁷ or within medical applications⁵⁸.

3.3.3 Federated Learning

Federated Learning (FL) evolves the Distributed Learning approach further, to leverage private data for the training of machine learning models while respecting the privacy of the data. This is a fundamental shift from more accepted methods and is shown in Figure XA above (right example).

In FL, distributed entities send their model updates to a Federated Learning server for aggregation. The steps are repeated in multiple rounds until an acceptable accuracy level is achieved. FL claims numerous advantages, such as data privacy, efficient bandwidth usage, and low latency. However, as a technology still nascent, there are doubts as to its suitability for all ML related applications and domains.

3.3.3.1 Data partitioning

One of the key decisions when using FL is to decide how the data is partitioned. Depending on the data distribution characteristics, data can be partitioned in a horizontal or vertical manner.

When the data located at the different clients is partitioned by samples, then each data set has the same features across all devices. i.e. each device learns to classify or predict the same features in the data. When using a vertical approach, the data is sampled in feature space. This means each device looks for a subset of specific features in order to train the local model. i.e. Each device learns only those features it is capable of deducing from the specific data types on each device.

This is perhaps shown more clearly in the following figure⁵⁹, where x_n is used to denote the n-th feature in the data and y is used to denote the correct label to apply.

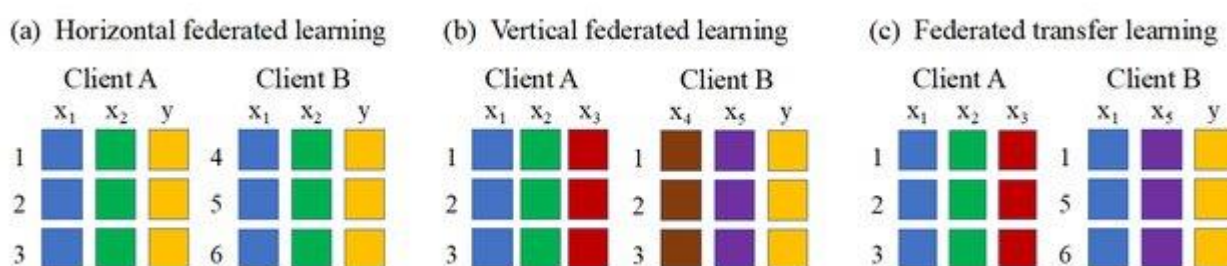


Figure 3. Categorization of Federated Learning

Federated Transfer Learning is introduced in the scenarios that the two data sets differ not only in samples but also in feature space with only a small portion of the feature space and sample space overlapped.

Where the same types of data are available at each client a Horizontal FL is the most appropriate. Where the data type differs between clients a Vertical FL is more appropriate. And finally, where the overlap in data types and samples is not great Transfer FL can be applied.

⁵⁷ Pacheco, A., Flores, E., Sánchez, R., & Almanza-García Salvador. (2018). Smart classrooms aided by deep neural networks inference on mobile devices. In 2018 IEEE International Conference Electro/Information Technology (EIT) (pp. 0605–0609). IEEE.

⁵⁸ Lee, K. H., & Verma, N. (2013). A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals. IEEE Journal of Solid-State Circuits, 48(7), 625–637.

⁵⁹ Chen, S., Xue, D., Chuai, G., Yang, Q., & Liu, qi. (2020). FL-QSAR: a federated learning based QSAR Prototype for collaborative drug discovery. ArXiv Preprint. <https://doi.org/10.1101/2020.02.27.950592>

3.3.3.2 Aggregation communication

Another key decision is the design of the aggregation mechanism. The aggregation mechanism can be designed in a centralised or distributed manner. Remember while the learning is distributed, the aggregation of the learned models may not be.

This decision also affects the associated communication overhead and eventual cost of the system. Centralised aggregators are more efficient in terms of communication usage, however they are more prone to failures due to their centralised nature. Konečný provides a more detailed examination of the relative communication costs of FL aggregation⁶⁰.

The following table summarises some FL approaches under these categories.

Table 2. Comparison among some Federated Learning implementations

Learning system	Horizontal data partition	Vertical data partition	Secure model aggregation	Communication design
SecureBoost ⁶¹	N	Y	N	Centralised
Tree-based FL ⁶²	Y	N	N	Decentralised
Ridge Regression Federated Learning ⁶³	Y	N	Y	Centralised
Federated Transfer Learning ⁶⁴ (Liu,	N	Y	N	Centralised
Federated Multi-Task Learning ⁶⁵	Y	N	Y	Centralised
Bayesian Nonparametric Federated Learning ⁶⁶	Y	N	Y	Centralised

3.3.3.3 Implementation challenges

FL has been experimentally applied in many application areas. However, the implementation of real systems can be tricky due to the following factors as identified by Li⁶⁷:

⁶⁰ Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. ArXiv Preprint ArXiv:1610.05492.

⁶¹ Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., & Yang, Q. (2021). Secureboost: lossless federated learning framework. IEEE Intelligent Systems, 36(6), 87–98.

⁶² Zhao, L., Ni, L., Hu, S., Chen, Y., Zhou, P., Xiao, F., & Wu, L. (2018). Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In IEEE INFOCOM 2018–IEEE Conference on Computer Communications (pp. 2087–2095). IEEE.

⁶³ Chen, Y.-R., Rezapour, A., & Tzeng, W.-G. (2018). Privacy-preserving ridge regression on distributed data. Information Sciences, 451, 34–49.

⁶⁴ Liu, Y., Kang, Y., Xing, C., Chen, T., & Yang, Q. (2020). A secure federated transfer learning framework. IEEE Intelligent Systems, 35(4), 70–82.

⁶⁵ Smith, V., Chiang, C.-K., Sanjabi, M., & Talwalkar, A. S. (2017). Federated multi-task learning. Advances in Neural Information Processing Systems, 30.

⁶⁶ Yurochkin, M., Agarwal, M., Ghosh, S., Greenwald, K., Hoang, N., & Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In International Conference on Machine Learning (pp. 7252–7261). PMLR.

⁶⁷ Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

3.3.3.3.1 Communication Bottlenecks

It is necessary to develop communication-efficient methods that iteratively send small messages or model updates as part of the training process. While this is better than sending the entire dataset over the network, there is still communication overhead.

Conceptually, this has to be addressed one of two ways: reducing the total number of messages required, and reducing the size of transmitted messages.

3.3.3.3.2 Systems heterogeneity

In federated settings, there is significant variability in the systems characteristics across the network, as devices may differ in terms of hardware, network connectivity, and battery power. Put simply this means that there is a variance in the time required to learn at each client node.

3.3.3.3.3 Trade-off between privacy and efficiency

Methods to enhance the privacy of federated learning have been suggested by using secure multiparty computation or differential privacy. However, these approaches provide privacy but at the cost of reduced model performance or system efficiency. For example, differential privacy requires the addition of noise to the data to enhance privacy protection. However, we sacrifice model accuracy in doing so, as the signal to noise ratio changes.

3.3.4 Conformal learning

Conformal prediction is a relatively new paradigm developed at the beginning of 2000, see⁶⁸ for an overview. It was originally developed for transductive setting⁶⁹ or online learning. The latter is efficient in terms of data usage but is also computationally expensive. Recent studies focus on *Inductive Conformal Prediction (ICP)*⁷⁰. *ICP* trains the learning model only once, however a part of the training dataset should be put aside for model calibration. Unless stated otherwise, in the rest of the text we will refer to Inductive Conformal Framework when talking about conformal prediction.

Conformal prediction^{71,72} is a framework that produces predictions with accuracy guarantees. For a given value of significance level $\epsilon \in (0,1)$, a conformal predictor is guaranteed to make exactly ϵ errors in the long run. This is achieved at a price of a reduction in prediction precision. Instead of predicting a single class label, in the case of classification, or a single number, in the case of regression, a conformal predictor outputs a range prediction, that is a set of class labels or an interval that contains the true value with probability $1 - \epsilon$. Construction of a conformal predictor with $\epsilon = 0$ is a trivial task. It is enough to output all class labels or an unbounded interval in the case of classification and regression respectively. However, such a predictor is of low value, that is, it is not *efficient*. The question thus is how to guarantee the given level of error rate ϵ by

⁶⁸ Linusson, Henrik. Nonconformity Measures and Ensemble Strategies: An Analysis of Conformal Predictor Efficiency and Validity. Diss. Department of Computer and Systems Sciences, Stockholm University, 2021.

⁶⁹ Vovk, Vladimir. Transductive conformal predictors. IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Berlin, Heidelberg, 2013.

⁷⁰ Papadopoulos, Harris. Inductive conformal prediction: Theory and application to neural networks. Rijeka: INTECH Open Access Publisher, 2008.

⁷¹ Shafer, Glenn, and Vladimir Vovk. A tutorial on conformal prediction. Journal of Machine Learning Research 9.3 (2008).

⁷² Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer Science & Business Media, 2005.

producing the smallest prediction regions. This property is achieved via the definition of a proper nonconformity function that succeeds in measuring the *strangeness* or *nonconformity* of every data instance. The difference between traditional Machine Learning (ML) and Conformal Learning frameworks is schematically presented in the next Figure. In the traditional ML the dataset is usually split into training and test sets. The training set is used to train a predictive ML model. The latter is afterwards tested on the held-out test set. Conformal Learning framework is a bit more complicated. As in the previous case, the dataset is split into training and test sets. The latter set is used as in the traditional ML. The training dataset, however, is further split into proper training and calibration sets. The proper training set is used to train a predictive model, and the calibration set is used for its calibration. The trained model and the user-defined nonconformity function are used to calculate calibration scores on the calibration set. These calibration scores are at the end used to generate region predictions that will produce no more than ϵ errors.

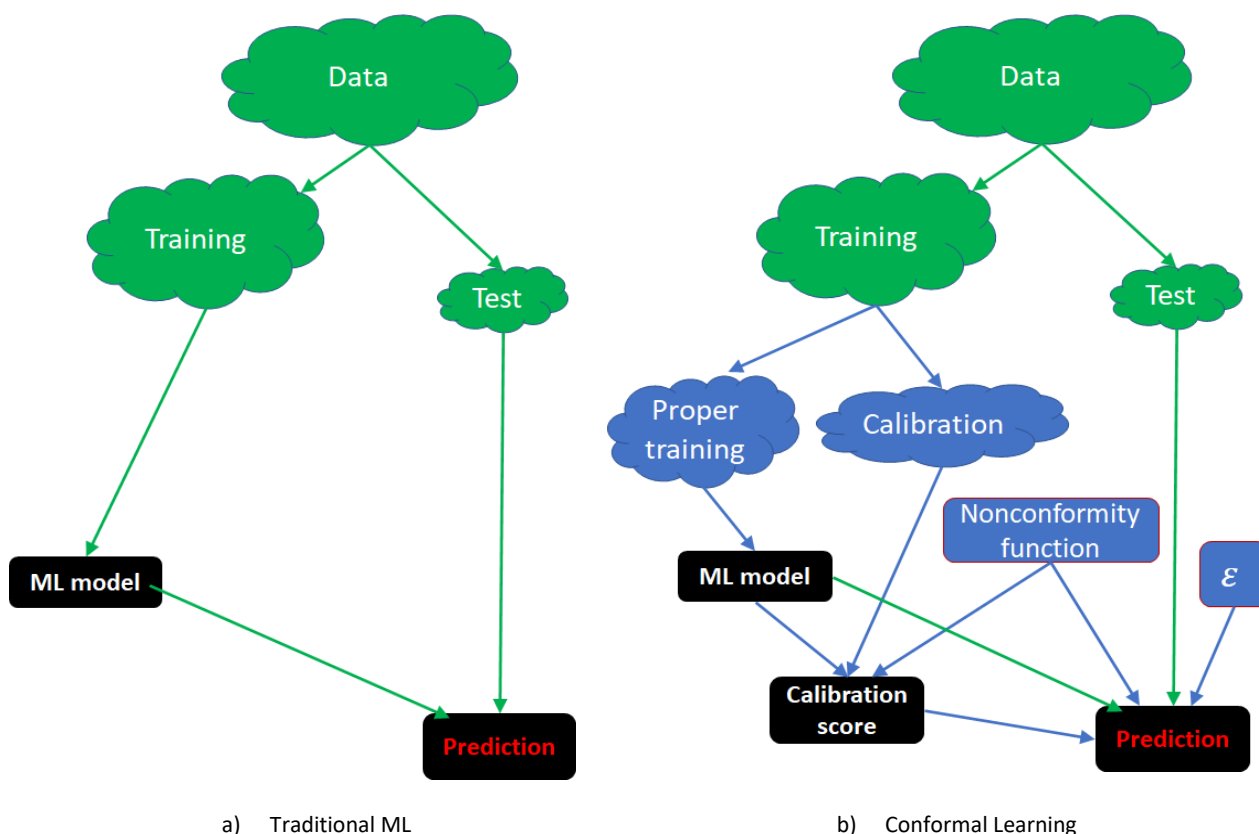


Figure 4. Traditional Machine Learning versus Conformal Learning

As we can see, the Conformal Learning framework has only 2 parameters: the nonconformity function and the allowed error rate ϵ . The value of ϵ is usually application specific. However, the choice of nonconformity function has a large impact on the efficiency and performance of a conformal predictor. Theoretically, any function, even a random function, can be used. But the proper choice of the nonconformity function will result in Conformal predictors with the desired properties.

3.3.4.1 Estimation of calibration scores

The core difference of Conformal learning as compared to traditional ML is estimation of calibration scores. For this, the values of the nonconformity function are calculated on every data point from the calibration set. The obtained values are then sorted and are used to estimate the empirical probability distribution function of calibration scores.

3.3.4.2 Classification

After calculating calibration scores and sorting them, a conformal classifier is ready to generate region prediction. For every datapoint from the test set, the calibration scores are calculated once again. This time, it is done for every possible class label. For example, if there are 2 classes, the calibration score will be calculated twice for every data point from the test set. These values are after that compared to the sorted calibration scores calculated for the calibration dataset with the aim to calculate the corresponding quantiles. Let us consider an example of the distribution of calibration scores shown with a green line in Fig.4. Assume that $\epsilon = 0.05$ and the values of nonconformity functions for class_1 and class_2 of a particular test datapoint are equal to 0.42 and 0.94, see red lines in the next Figure. From this distribution, we can estimate what is the fraction of elements from the calibration data set that had larger values of nonconformity score. For class_1, it is 43.03%. It means that 43.03% of instances from the calibration dataset were more “non-conformal” or stranger then the current datapoint associated with class_1. Given that $0.4303 > \epsilon$, this class label should be included in the final prediction set for the current datapoint. As for class_2, only 3.33% of instances from the calibration dataset had larger values of nonconformity score. As $0.333 < \epsilon$, the probability of making a mistake by not including this label into the final prediction set is less than the allowed error rate of 5% and thus this label should not be included in the final region prediction. The detailed calculations for this example can be found in the relevant repository⁷³.

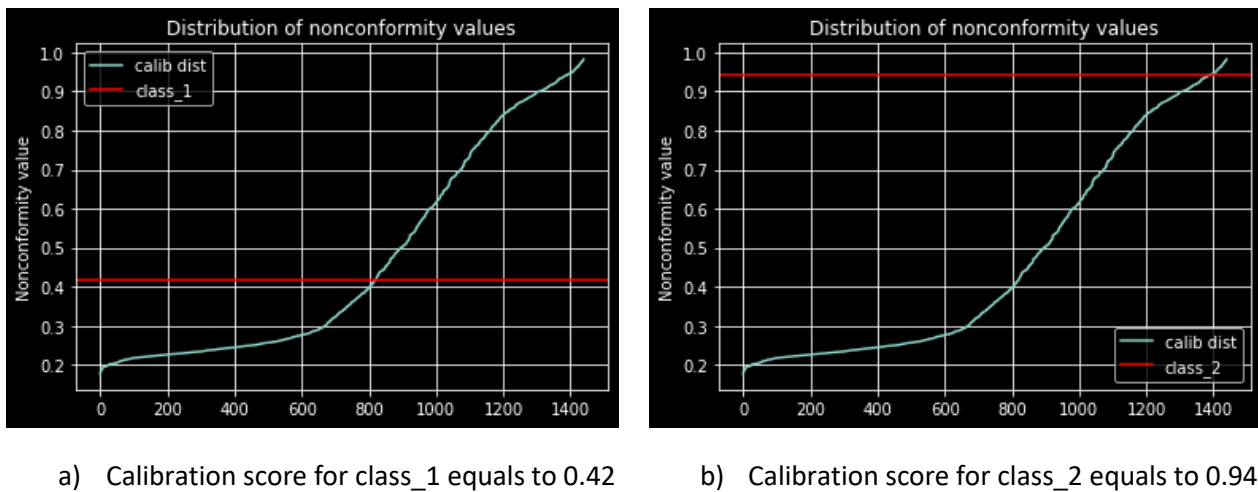


Figure 5. Calculation of calibration scores for classification

3.3.4.3 Regression

To understand how the conformal regression works, we will discuss the following example. Let nonconformity function be defined as the absolute difference between the predicted value $\hat{y}_i$ and the true value y_i $y_i - \hat{y}_i \vee$ (this function is known as Absolute Error function). Then, we can calculate this score for every point in the calibration set and estimate its distribution. From this distribution, we can take the quantile corresponding to the required level of ϵ , let it be α_ϵ . From here, we have $y_i - \hat{y}_i \vee \alpha_\epsilon$ and $\hat{y}_i - \alpha_\epsilon \leq y_i \leq \hat{y}_i + \alpha_\epsilon$. This means that the interval $[\hat{y}_i - \alpha_\epsilon; \hat{y}_i + \alpha_\epsilon]$ will contain the true value with the probability $1 - \epsilon$. Of course, the Absolute Error function is not the only possible choice of nonconformity function for regression problems. Another common solution is Signed error function defined as $y_i - \hat{y}_i$. In this case, the upper and the lower bounds of the regression interval are identified from separate distributions.

⁷³ https://github.com/marharyta-aleksandrova/conformal-learning/blob/main/theory/conformal_learning_from_scratch.ipynb



From the construction procedure, we can see that, as opposed to conformal classifiers, a conformal regressor has the same length of the prediction interval for every point in the test set and in the production. For our example, it is $2\alpha_\epsilon$. This approach is less flexible and it does not adjust to the changes in the data distribution. Normalised conformal regression⁷⁴ is the technique that can generate prediction intervals of different sizes depending on the difficulty of every datapoint.

3.2.6.1 Measuring efficiency and Impact of nonconformity functions

By definition, all conformal predictors with the same value of ϵ will have the same accuracy. Thereby, other metrics are used to compare the efficiency of different conformal predictors.

The efficiency of a conformal predictor depends on the choice of nonconformity function. There are two groups of nonconformity functions: model-agnostic and model-dependent. Model-dependent nonconformity functions are defined based on the underlying prediction model. Such functions can depend on the distance to the separating hyperplane in SVM⁷⁵, or the distance between instances in KNN classifier⁷⁶. These nonconformity functions are model-specific, thereby, one does not draw generalised conclusions about their behaviour. Model-agnostic functions are usually based on the output of the predictive model and can be applied to construct conformal predictors based on various underlying ML models. In a recent study it was shown that model-agnostic nonconformity functions do have some general characteristics.

3.3.4.4 Classification

In the case of classification, the *efficiency* of a conformal predictor is often measured in terms of 2 metrics: *avgC*, which stands for the average number of predicted class labels per instance, and *oneC*, which stands for the fraction of produced singleton predictions. Naturally, one would want to minimise *avgC* and maximise *oneC* at the same time. A recent study⁷⁷ showed that *the usage of margin as a nonconformity function results in higher oneC and the usage of inverse probability (also known as hinge⁷⁸) results in lower values of avgC*. Inverse probability nonconformity function is defined by the equation $\Delta[h(x_i, y_i)] = 1 - P_h(y_i \vee x_i)$, where x_i is the analysed data instance, y_i is a tentative class label, and $P_h(y_i \vee x_i)$ is the probability assigned to this label given the instance x_i by the underlying classifier h . Margin is defined by the following formula $\Delta[h(x_i, y_i)] = \max_{y \neq y_i} P_h(y \vee x_i) - P_h(y_i \vee x_i)$, and it measures how different is the probability of the label y_i from the probability of another most probable class label. The authors used 21 datasets to demonstrate the statistical significance of this relationship. However, this was done for the case where the baseline classifier was either a single artificial neural network (ANN) or an ensemble of bagged ANNs.

Combining characteristics of both *margin* and *inverse probability* nonconformity functions is a tempting idea. In recent years many authors have dedicated their efforts to understand how one can generate more efficient

⁷⁴ Papadopoulos, Harris, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008). 2008.

⁷⁵ Balasubramanian, Vineeth Nallure, et al. Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. 2009 36th Annual Computers in Cardiology Conference (CinC). IEEE, 2009.

⁷⁶ Proedrou, Kostas, et al. Transductive confidence machines for pattern recognition. European Conference on Machine Learning. Springer, Berlin, Heidelberg, 2002.

⁷⁷ Johansson, Ulf, et al. Model-agnostic nonconformity functions for conformal classification. 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

⁷⁸ In the rest of the text, we will refer to this relationship as a baseline or original pattern (relationship)

conformal predictions through a combination of several conformal predictors. Papers^{79,80} studied how to aggregate conformal predictions based on different training algorithms. Various strategies were proposed for such combination: via p -values⁸¹, a combination of monotonic conformity scores⁸², majority voting⁸³, out-of-bag calibration⁸⁴, or via established results in classical statistical hypothesis testing⁸⁵. The challenge of every combination of conformal predictors is to retain *validity*, that is to achieve the empirical error rate not exceeding the predefined value ϵ . This property is usually demonstrated experimentally and some authors provide guidelines on which values of significance levels should be used for individual conformal algorithms to achieve the desired validity of the resulting combination.

Some authors⁸⁶ have extended the previous study with the following contributions:

1. It was shown that although the previously observed pattern (*the usage of margin as a nonconformity function results in higher oneC and the usage of inverse probability results in lower values of avgC*) does hold in the majority of the cases, the choice of the best nonconformity function can depend on the analysed dataset and the chosen underlying classification model. For example, k -nearest neighbours classifier performs best with *margin*. *Margin* is also the best choice in case of some datasets regardless of the chosen classification model.
2. The authors proposed a method to combine both nonconformity functions. Their experimental evaluation shows that this combination always results in better or the same performance as *inverse probability*, thus allowing to increase the value of *oneC* and decrease the value of *avgC*. In some cases, the proposed combination outperforms both *inverse probability* and *margin* in terms of both efficiency metrics.
3. Authors discuss several aspects of how the accuracy of the baseline classifier can impact the performance of the resulting conformal predictor. In particular, if the baseline prediction accuracy is very good, then nonconformity function has no impact on the efficiency. Also, the accuracy of the baseline classifier strongly correlates with the fraction of singleton predictions that contain the true label. In this way, the accuracy can be an indicator of the usefulness of the *oneC* metric.

There are also other useful metrics⁸⁷ that can be used to measure the efficiency of conformal classifiers, such as *zeroC* (number of empty predictions) *oneAcc* (accuracy of singleton predictions) etc. Depending on the application, a certain metric can be preferred. In the future work, it would be interesting to further study

⁷⁹ Yang, Yachong, and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. arXiv preprint arXiv:2104.13871 (2021).

⁸⁰ Toccaceli, Paolo, and Alexander Gammerman. Combination of inductive mondrian conformal predictors. Machine Learning 108.3 (2019): 489-510.

⁸¹ Toccaceli, Paolo, and Alexander Gammerman. Combination of conformal predictors for classification. Conformal and Probabilistic Prediction and Applications. PMLR, 2017.

⁸² Gauraha, Niharika, and Ola Spjuth. Synergy conformal prediction. Conformal and Probabilistic Prediction and Applications. PMLR, 2021.

⁸³ Cherubin, Giovanni. Majority vote ensembles of conformal predictors. Machine Learning 108.3 (2019): 475-488.

⁸⁴ Linusson, Henrik, Ulf Johansson, and Henrik Boström. Efficient conformal predictor ensembles. Neurocomputing 397 (2020): 266-278.

⁸⁵ Toccaceli, Paolo. Conformal predictor combination using neyman–pearson lemma. Conformal and Probabilistic Prediction and Applications. PMLR, 2019.

⁸⁶ Aleksandrova, Marharyta, and Oleg Chertov. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. Conformal and Probabilistic Prediction and Applications. PMLR, 2021.

⁸⁷ Johansson, Ulf, Henrik Boström, and Tuve Löfström. Conformal prediction using decision trees. 2013 IEEE 13th international conference on data mining. IEEE, 2013.

how the choice of nonconformity function can shape the properties of conformal classifiers with the aim to generate relevant recommendations.

3.3.4.5 Regression

Regression, being a continuous prediction task, results in interval predictions. Measuring efficiency in this case means measuring the size of the prediction interval. As it was discussed above, in the case of traditional conformal regression, the size of the prediction interval stays the same for the same model. However, for normalised conformal regression it changes. In this case it might make sense to consider various metrics, such as average, minimum or median size of the prediction interval. Among 2 oftenly used nonconformity functions for regression, namely Absolute Error and Signed Error, the first one is known to result in more efficient conformal predictors both in traditional and normalised settings. However, further research is required to better understand the impact of nonconformity functions on the efficiency of conformal regressors.

3.3.4.6 Applications of Conformal Learning

Apart from the obvious application of conformal learning framework for generating predictions with accuracy guarantees, conformal learning can be also applied to solve other problems. We can mention the following examples: anomaly detection^{88,89,90}, statistical hypothesis testing⁹¹, clustering^{92,93}, multi-label classification⁹⁴ etc. This theory was also successfully applied in various practical tasks, such as recommender systems^{95,96}, internet security⁹⁷, medicine⁹⁸ and others.

⁸⁸ Laxhammar, Rikard, and Göran Falkman. Sequential conformal anomaly detection in trajectories based on Hausdorff distance. 14th International Conference on Information Fusion. IEEE, 2011.

⁸⁹ Laxhammar, Rikard. Conformal anomaly detection: Detecting abnormal trajectories in surveillance applications. Diss. University of Skövde, 2014.

⁹⁰ Bates, Stephen, et al. "Testing for outliers with conformal p-values." arXiv preprint arXiv:2104.08279 (2021).

⁹¹ Vovk, Vladimir. "Conformal testing in a binary model situation." In Conformal and Probabilistic Prediction and Applications, pp. 131-150. PMLR, 2021.

⁹² Nourtdinov, Ilia, James Gammerman, Matteo Fontana, and Daljit Rehal. Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. *Neurocomputing* 397 (2020): 279-291.

⁹³ Cherubin, Giovanni, Ilia Nourtdinov, Alexander Gammerman, Roberto Jordaney, Zhi Wang, Davide Papini, and Lorenzo Cavallaro. "Conformal clustering and its application to botnet traffic." In International Symposium on Statistical Learning and Data Sciences, pp. 313-322. Springer, Cham, 2015.

⁹⁴ Papadopoulos, Harris. "A cross-conformal predictor for multi-label classification." In IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 241-250. Springer, Berlin, Heidelberg, 2014.

⁹⁵ Himabindu, Tadiparthi VR, Vineet Padmanabhan, and Arun K. Pujari. Conformal matrix factorization based recommender system. *Information Sciences* 467 (2018): 685-707.

⁹⁶ Kagita, Venkateswara Rao, Arun K. Pujari, Vineet Padmanabhan, Sandeep Kumar Sahu, and Vikas Kumar. "Conformal recommender system." *Information Sciences* 405 (2017): 157-174.

⁹⁷ Liu, Chunbo, Mengmeng Liang, Jingwen Hou, Zhaojun Gu, and Zhi Wang. LogCAD: An Efficient and Robust Model for Log-Based Conformal Anomaly Detection. *Security and Communication Networks* 2022 (2022).

⁹⁸ Liu, Li, Xianghao Zhan, Rumeng Wu, Xiaoqing Guan, Zhan Wang, Wei Zhang, Mert Pilanci, You Wang, Zhiyuan Luo, and Guang Li. Boost AI Power: Data Augmentation Strategies With Unlabeled Data and Conformal Prediction, a Case in Alternative Herbal Medicine Discrimination With Electronic Nose. *IEEE Sensors Journal* 21, no. 20 (2021): 22995-23005.

3.3.5 Causal analysis

Thanks to the technological and computational advances during the last decades, scientists were able to successfully tackle non-trivial problems from different research areas, with causality being a prominent example. One of the fundamental problems of causality theory is to determine the causal relationship between two or more variables. This problem is known as *causal discovery*, *causal identification* or *structure learning*^{99,100}. For example, given altitude and temperature, we want to answer the question if the temperature has an effect on altitude, or if altitude has an effect on temperature. This is of particular interest since if such a causal relationship is known then one can predict the effects on a system in case of an intervention or a perturbation.

Controlled experimentation, or A/B tests, are considered to be a golden standard for causal discovery^{101, 102}. In such experiments, there are two identical groups with only one variation. The only variable that is varied (intervened on) is the potential cause. This procedure allows estimating the causal effect of this variable in a given system. A/B tests are widely used in practical applications. For example, testing the efficacy of medications is usually done with A/B tests, see¹⁰³ for an example. In this case, the first group, also known as the *control group*, receives no medication or a placebo, and the second group, known as the *intervention group*, receives the real medication. The results show the true effect (if any) of the medication on human health. However, such tests are often too expensive, unethical, or even technically impossible to execute. For example, to test the effect of smoking on health with this approach, one needs two non-smoker groups. Next, the members of one group should be forced to smoke, and the others not do so. Therefore, it is of great interest to determine causal relationships from observational data only.

There exist many methods which are able to determine causal relationships from observational data. The basic idea of structure learning emerged from¹⁰⁴ as *path analysis*. Judea Pearl presented in his work¹⁰⁵ a comprehensive theory of causality and unified the probabilistic, manipulative, counterfactual, and structural approaches to causation. From this work we have the following key point. If there is a statistical association, e.g. two variables X and Y are dependent, then one of the following is true: 1) there is a causal relationship, either X has an effect on Y or Y has an effect on X ; 2) there is a common cause (*confounder*) that has an effect on both X and Y ; 3) there is a possibly unobserved common effect of X and Y that is conditioned upon data acquisition (selection bias); or 4) there can be a combination of these. From there on, a lot of research has been conducted to develop theoretical approaches and methods for structure learning.

⁹⁹ Spirtes, Peter, et al. Causation, prediction, and search. MIT press, 2000.

¹⁰⁰ Neuberger, Leland Gerson. "Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000." *Econometric Theory* 19.4 (2003): 675-685.

¹⁰¹ Kohavi, Ron, and Roger Longbotham. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7.8 (2017): 922-929.

¹⁰² Young, Scott WH. Improving library user experience with A/B testing: Principles and process. *Weave: Journal of Library User Experience* 1.1 (2014).

¹⁰³ Thase, Michael E., et al. Impact of pharmacogenomics on clinical outcomes for patients taking medications with gene-drug interactions in a randomized controlled trial. *The Journal of clinical psychiatry* 80.6 (2019): 22039.

¹⁰⁴ Verhulst, Brad, Lindon J. Eaves, and Peter K. Hatemi. Correlation not causation: The relationship between personality traits and political ideologies. *American journal of political science* 56.1 (2012): 34-51.

¹⁰⁵ Pearl, Judea. *Models, reasoning and inference*. Cambridge, UK: CambridgeUniversityPress 19 (2000): 2.

In general, all methods for structure learning exploit the complexity of the marginal and conditional probability distributions in some way ^{106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129}.

-
- ¹⁰⁶ Rebane, George, and Judea Pearl. The recovery of causal poly-trees from statistical data. arXiv preprint arXiv:1304.2736 (2013). <http://arxiv.org/abs/1304.2736>
- ¹⁰⁷ Spirtes, Peter, et al. Causation, prediction, and search. MIT press, 2000.
- ¹⁰⁸ Janzing, Dominik, et al. Information-geometric approach to inferring causal directions. Artificial Intelligence 182 (2012): 1-31.
- ¹⁰⁹ Sgouritsa, Eleni, et al. Inference of cause and effect with unsupervised inverse regression. Artificial intelligence and statistics. PMLR, 2015.
- ¹¹⁰ Friedman, Nir, and Iftach Nachman. Gaussian process networks. arXiv preprint arXiv:1301.3857 (2013). <http://arxiv.org/abs/1301.3857>
- ¹¹¹ Kano, Yutaka, and Shohei Shimizu. Causal inference using nonnormality. Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion. 2003.
- ¹¹² Shimizu, Shohei, et al. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research 7.10 (2006).
- ¹¹³ Shimizu, Shohei, Aapo Hyvarinen, and Yoshinobu Kawahara. A direct method for estimating a causal ordering in a linear non-gaussian acyclic model. arXiv preprint arXiv:1408.2038 (2014). <http://arxiv.org/abs/1408.2038>
- ¹¹⁴ Sun, Xiaohai, Dominik Janzing, and Bernhard Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. Ninth International Symposium on Artificial Intelligence and Mathematics (AIMath 2006). 2006.
- ¹¹⁵ Sun, Xiaohai, Dominik Janzing, and Bernhard Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. Neurocomputing 71.7-9 (2008): 1248-1256.
- ¹¹⁶ Hoyer, Patrik, et al. Nonlinear causal discovery with additive noise models. Advances in neural information processing systems 21 (2008).
- ¹¹⁷ Hoyer, Patrik, et al. Nonlinear causal discovery with additive noise models. Advances in neural information processing systems 21 (2008). <http://arxiv.org/abs/0909.4386>
- ¹¹⁸ Mooij, Joris, et al. Regression by dependence minimization and its application to causal inference in additive noise models. Proceedings of the 26th annual international conference on machine learning. 2009.
- ¹¹⁹ Stegle, Oliver, et al. Probabilistic latent variable models for distinguishing between cause and effect. Advances in neural information processing systems 23 (2010).
- ¹²⁰ Mooij, Joris M., et al. On causal discovery with cyclic additive noise models. Advances in neural information processing systems 24 (2011).
- ¹²¹ Zhang, Kun, and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. arXiv preprint arXiv:1205.2599 (2012). <http://arxiv.org/abs/1205.2599>
- ¹²² Danusis, Povilas, et al. Inferring deterministic causal relations. arXiv preprint arXiv:1203.3475 (2012). <http://arxiv.org/abs/1203.3475>
- ¹²³ Hyvärinen, Aapo, and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. Journal of Machine Learning Research 14.Jan (2013): 111-152.
- ¹²⁴ Shimizu, Shohei. LiNGAM: Non-Gaussian methods for estimating causal structures. Behaviormetrika 41.1 (2014): 65-98.
- ¹²⁵ Peters, Jonas, and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. Biometrika 101.1 (2014): 219-228. <https://doi.org/10.1093/biomet/ast043>
- ¹²⁶ Peters, Jonas, et al. Causal discovery with continuous additive noise models. (2014).
- ¹²⁷ Nowzohour, Christopher, and Peter Bühlmann. Score-based causal learning in additive noise models. Statistics 50.3 (2016): 471-485.
- ¹²⁸ Park, Gunwoong, and Youngwhan Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. Journal of the Korean Statistical Society 49.1 (2020): 276-292. <http://arxiv.org/abs/1901.10134>
- ¹²⁹ Chen, Wenyu, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. Biometrika 106.4 (2019): 973-980.



Under certain assumptions, these methods are then able to solve the task of causal discovery. Let C denote the cause and E the effect. Then their joint density can be expressed with $p_{C,E}(c, e)$. This joint density can be factorised into either (1) $p_C(c) \cdot P_{E|C}(e \vee c)$ or (2) $p_E(e) \cdot P_{C|E}(c \vee e)$. The idea is then that (1) gives models of lower total complexity than (2) and this allows us to conclude the causal relationship direction. Intuitively, this makes sense, because the effect contains information from the cause but not vice-versa (of course, under the assumption that there are no cycles aka feedback loops). Therefore, (2) has at least as much complexity as (1). However, the definition of complexity is ambiguous. For example, one can say that " p_C contains no information about $P_{E|C}(e|c)$ " and then draw partial conclusions about the causal direction in a given system. This complexity question is often colloquially referred to as *breaking the symmetry*, that is $p_C(c) \cdot P_{E|C}(e \vee c) \neq P_E(e) \cdot P_{C|E}(c \vee e)$.

One particular group of causal discovery methods is based on *Additive Noise Models* (ANMs). These methods, as the name suggests, exploit the additivity of the random hidden noise. ANMs received a lot of attention as they are well established and yielded many good results¹³⁰. Silva et al. introduced in ¹³¹ a method for learning the structure of linear latent variable models. The main assumption in their work is that each variable is a linear function of its parents plus an additive error term of positive finite variance. Hoyer et al. generalised the linear framework of additive noise models to the nonlinear case¹³². Earlier works often assumed linear models for continuous variables. The authors showed that if data contains non-Gaussian variables, then this can help in distinguishing the causal directions and identifying the causal graph. Mooij et al. introduced Resit method¹³³. This method is based on the idea of minimising the statistical dependence between the regressors and residuals¹³⁴. The authors demonstrated that if the residuals are no longer dependent on the input, then regression can successfully model the causal dependence. This method does not need to assume a particular distribution of the noise because any form of regression can be used (e.g., Linear Regression), and it is well suited for the task of causal inference in additive noise models. Next, Mooij et al. introduced a method to determine the causal relationship in cyclic additive noise models and showed that such models are generally identifiable in the bivariate, Gaussian-noise case¹³⁵. Their method works for continuous data and can be seen as a special case of nonlinear independent component analysis. Later, Peters and Bühlmann proved in ¹³⁶ *full identifiability*¹³⁷ of linear Gaussian structural equation models if all the noise variables have the same variance. In the next work, Peters et al. proposed a method that can identify the directed acyclic graph from the distribution under mild conditions¹³⁸. In contrast, previous methods assumed faithfulness and could only

¹³⁰ Kpotufe, Samory, et al. Consistency of causal inference under the additive noise model. International Conference on Machine Learning. PMLR, 2014.

¹³¹ Silva, Ricardo, et al. Learning the Structure of Linear Latent Variable Models. Journal of Machine Learning Research 7.2 (2006).

¹³² Hoyer, Patrik, et al. Nonlinear causal discovery with additive noise models. Advances in neural information processing systems 21 (2008).

¹³³ Mooij, Joris, et al. Regression by dependence minimization and its application to causal inference in additive noise models. Proceedings of the 26th annual international conference on machine learning, 2009.

¹³⁴ The residuals are defined as the difference between the actual output and the predicted output.

¹³⁵ Mooij, Joris M., et al. On causal discovery with cyclic additive noise models. Advances in neural information processing systems 24 (2011).

¹³⁶ Peters, Jonas, and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. Biometrika 101.1 (2014): 219-228.

¹³⁷ Full identifiability means that not only the skeleton of the causal graph is recoverable but also the arrows are

¹³⁸ Peters, Jonas, et al. Causal discovery with continuous additive noise models. (2014).

identify the Markov equivalence class of the graph¹³⁹. Finally, the authors of^{140,141} proved that linear Gaussian models with different error variance can be also identifiable. In their method, referred to as Uncertainty Scoring, this is done by ordering variables according to the law of total variances and then performing independence tests between them. Park extended this result to additive noise models in¹⁴².

As we can see, many researchers contributed to the development of ANMs-based causal discovery methods and widened our understanding of their application cases. However, no previous research work analysed how the level of noise variance relative to that of the cause variance can impact the accuracy of these methods. In the real world, it can occur that noise levels change drastically from cause to effect. It can happen, for example, if the data collection process has a lot of interference like in outer space.

3.3.6 Distributed Compression & Compressed Learning

Communication and data movement are often the most expensive parts during the data collection. As a result, the problem of data compression has been studied, with a view to both reducing the amount of data transmitted, and its final storage. Whilst many compression algorithms exist, they can be classified into two broad categories. Examining each method in turn.

3.3.6.1 Sparsification

It allows an optimal reconstruction of the actual signal by using significantly fewer samples than required by the Nyquist criteria¹⁴³ (Landau, 1967). Several studies have shown the benefits of what is termed "compressive sensing" in power consumption optimization for mobile¹⁴⁴ or wearable devices¹⁴⁵. Compressive sensing is a signal acquisition and reconstruction technique where signal sparsity or quantisation are exploited to achieve efficiency in energy consumption, bandwidth, and performance¹⁴⁶. Many IoT applications rely on sparse signals, or signals that do not change often. Therefore, applying fixed sensing intervals might prove to be inefficient for an application involving sparse signals. An example is where data is mostly zero, only passing on non-zero values for analysis. However, the downside is that this technique can only be used for sparse signals (data). However, compressive sensing has also gained attention in the cognitive radio communication field¹⁴⁷.

¹³⁹ Markov equivalence class refers to the class of graphs in which all graphs have the same skeleton

¹⁴⁰ Park, Gunwoong, and Youngwhan Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society* 49.1 (2020): 276-292. <http://arxiv.org/abs/1901.10134>

¹⁴¹ Chen, Wenyu, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika* 106.4 (2019): 973-980.

¹⁴² Park, Gunwoong. Identifiability of Additive Noise Models Using Conditional Variances. *J. Mach. Learn. Res.* 21.75 (2020): 1-34.

¹⁴³ Landau, H. J. (1967). Sampling, data transmission, and the Nyquist rate. *Proceedings of the IEEE*, 55(10), 1701–1706.

¹⁴⁴ Fragkiadakis, A., Charalampidis, P., & Tragos, E. (2014). Adaptive compressive sensing for energy efficient smart objects in IoT applications. 2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 1–5.

¹⁴⁵ Bellasi, D. E., & Benini, L. (2015). Energy-efficiency analysis of analog and digital compressive sensing in wireless sensors. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 62(11), 2718–2729.

¹⁴⁶ Padalkar, S., Korlekar, A., & Pacharaney, U. (2016). Data gathering in wireless sensor network for energy efficiency with and without compressive sensing at sensor node. 2016 International Conference on Communication and Signal Processing (ICCCSP), 1356–1359.

¹⁴⁷ Sharma, S. K., Lagunas, E., Chatzinotas, S., & Ottersten, B. (2016). Application of compressive sensing in cognitive radio communications: A survey. *IEEE Communications Surveys & Tutorials*, 18(3), 1838–1860.

A slight variant of this is to apply an "adaptive compressive sensing" approach, where the sampling period varies depending on the context. When applied correctly, adaptive compressed sensing proves to be very effective where sampling rates are varied dynamically as and when required to conserve energy¹⁴⁸.

3.3.6.2 Quantisation

A second technique is to apply quantisation to the data. In contrast to the above approach, quantisation relies on reducing the precision of the data recorded. If the value to be measured is captured in a 32-bit float number, quantisation attempts to capture these values with fewer bits, 16 or even 4 bits. For example, consider a fridge storing milk, the temperature range is -4 to +12 for normal operation. Thus a four bit representation can capture all possible operational values. This reduces the communication and storage costs.

In contrast, to the sparsification approach this technique can be applied when applications require higher sampling rates. Thus, this method can be thought of as a more general approach, where it can be applied to individual data samples.

3.3.6.3 Compressed Learning

The advances in ML have initiated a transition from big-data to "big knowledge"¹⁴⁹. Pre-trained deep learning models allow the efficient analysis of big data such as the vast material in text, images and videos. "Big knowledge" implies knowledge is stored in a series of deep learning models, which are often very large themselves.

Model compression has emerged as an important area of research for deploying deep learning models on Internet-of-Things (IoT)¹⁵⁰. It is important to note that compression techniques (including some of those identified above) can also be applied to the learned model. This aids in the efficient transmission of deep learning models (both single model compression and compression of multiple models). Some work has optimised the delay when using these compressed models¹⁵¹.

However, for extremely memory-constrained scenarios, even the compressed models cannot fit within the memory of a single device and, as a result, must be distributed across multiple devices¹⁵². For RNN deep learning approaches efforts have been made to reduce the size of models when using resource

¹⁴⁸ Xu, K., Li, Y., & Ren, F. (2017). A data-driven compressive sensing framework tailored for energy-efficient wearable sensing. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 861–865.

¹⁴⁹ Chen, Z., Wang, S., Wu, D. O., Huang, T., & Duan, L.-Y. (2018). From data to knowledge: Deep learning model compression, transmission and communication. Proceedings of the 26th ACM International Conference on Multimedia, 1625–1633.

¹⁵⁰ Thakker, U., Beu, J., Gope, D., Dasika, G., & Mattina, M. (2019). Run-time efficient RNN compression for inference on edge devices. 2019 2nd Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2), 26–30.

¹⁵¹ Abdelmoniem, A. M., & Canini, M. (2021). DC2: Delay-aware Compression Control for Distributed Machine Learning. IEEE INFOCOM 2021-IEEE Conference on Computer Communications, 1–10.

¹⁵² Chen, Z., Wang, S., Wu, D. O., Huang, T., & Duan, L.-Y. (2018). From data to knowledge: Deep learning model compression, transmission and communication. Proceedings of the 26th ACM International Conference on Multimedia, 1625–1633.

Project 'cities2030' | H2020 ID | 101000640 | 'Co-creating resilient and sustainable food systems towards FOOD2030' | www.cities2030.eu
constrained devices and examining how much compression can be achieved ¹⁵³. Others have examined only the memory management aspect of applying CNN in resource constrained device ¹⁵⁴.

3.3.7 Summary

The following table attempts to summarise the previous approaches in terms of the challenges they address.

Table 3. Learning approaches and challenges

Approach	Connectivity	Bandwidth / Latency	Privacy	Transparency / fairness
"Classic" Learning	No	No	No	No
Distributed Learning	Partial	Partial	No	Yes
Federated Learning	Partial	Yes	Yes	Yes
Conformal Learning	No	No	No	Yes
Compressed Learning	No	Partial	Yes	No

3.4 Previous projects

Finally, in this third subsection, previous projects whose results are relevant for Cities2030 projects are analyzed and presented.

3.4.1 BUSPrime

Business Privacy Modelling and Execution (BusPRIME) is based around the idea of including privacy and data protection concerns in the business process design and execution. Recent incidents that involve customer data have raised the awareness of the privacy concerns and maintenance of customer trust is now becoming a challenge for businesses throughout the world. One of the principal methods to combat this, is to include a privacy-by-design¹⁵⁵ into the design of and hence the implementation of new or existing services. BusPRIME takes a rather unique approach to privacy-by-design; by including data-privacy aspects into the business process, as a means of ensuring data privacy compliance during specific stakeholder interactions.

¹⁵³ Thakker, U., Fedorov, I., Beu, J., Gope, D., Zhou, C., Dasika, G., & Mattina, M. (2019). Pushing the limits of rnn compression. 2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS), 18–21.

¹⁵⁴ Patel, S., Liu, T., & Guan, H. (2021). FreeLunch: Compression-based GPU Memory Management for Convolutional Neural Networks. 2021 IEEE/ACM Workshop on Memory Centric High Performance Computing (MCHPC), 1–8.

¹⁵⁵ Langheinrich, M. (2001). Privacy by design—principles of privacy-aware ubiquitous systems. International Conference on Ubiquitous Computing, 273–291.

Defining business processes that are privacy aware is a difficult task and requires the inclusion of security experts together with legal experts in the modelling process. There was a gap in the availability of tools that provide privacy and data protection regulatory assurance to the business process modeller at the design stage and also in the execution of those processes at run time.

The overarching project goal of BusPRIME is to provide organisations with a solution to model business processes that cater for privacy and data protection requirements and regulatory obligations at design time. This privacy by design approach is achieved by extending the BPMN 2.0 (Business Process Modelling and Notation) specification¹⁵⁶ with a set of privacy primitives and implementing a reference solution for integration with one of the already established vendors as well as providing a implementation for one of the popular open source BPM (Business Process Management) solutions¹⁵⁷.

3.4.1.1 Integration with BPMN

BPMN was chosen as it is a capable standards based approach to representing a business process and commercial tooling exists to manipulate it. Fundamentally, there are four data focused objects types defined in BPMN:

- Data objects capture documents used in a process, both in physical and digital form. Data objects are only available for the lifetime of the process;
- Data stores are a means to more permanently store data objects.
- Data Inputs allow you to show what kind of data is needed by Processes, Sub-Processes and Tasks in order to execute, so they are the data requirements.
- Data Outputs show data produced by Processes.

3.4.1.2 Extending BPMN

While BPMN is a generic language it allows "extensions" to the core model. It defines a BPMN extension mechanism, that allows software projects to extend the standard BPMN elements with additional attributes and tags based on a simple XML grammar. This flexibility is so useful that its often extended by researchers¹⁵⁸, either for dealing with processes of specific domains or for improving the language itself. This mechanism is supported by modelling tools to add non-standard elements or artefacts to satisfy a specific need and still have valid BPMN Core. The following example shows how the BPMN standard can be extended with custom attributes and tags:

```
<bpmn2:task id="Task_109" custom:category="marketing" name="Approve the Request">
  <bpmn2:extensionElements>
    <custom:item name="email" type="xs:string">
      <custom:value>approval@organisation.com</custom:value>
    </custom:item>
  </bpmn2:extensionElements>
</bpmn2:incoming>SequenceFlow_7</bpmn2:incoming>
```

¹⁵⁶ Business Process Model and Notation. (2014). Object Management Group. <https://www.omg.org/spec/BPMN/>

¹⁵⁷ BPMN Implementer's. (2022). Object Management Group. <https://www.bpmn.org/#tabs-implementers>

¹⁵⁸ Zarour, K., Benmerzoug, D., Guermouche, N., & Drira, K. (2019). A systematic literature review on BPMN extensions. Business Process Management Journal.

```
<bpmn2:outgoing>SequenceFlow_9</bpmn2:outgoing>
</bpmn2:task>
```

Figure 6. Sample BPMN extension definition

In the above example, an additional "email" attribute is added to "Task_100", that captures the situation where a marketing process has to be approved by "approval@organisation.com".

By implementing an extension to BPMN, data privacy aspects can be annotated to the Data Objects directly. Data objects are annotated with privacy constraints that restrict the use of the corresponding data objects. Privacy analysts can then see where a specific data object is being used (in an overall business process) and business analysts can see the restrictions on the data usage. The business process becomes the interaction space between experts, and thus allows organisational accountability¹⁵⁹.

3.4.1.3 Analysis of the approach

The approach taken by BusPRIME is very useful. BPMN is fundamentally a business process language and ideally suited to capturing a wide variety of processes, and the role of actors involved in those processes. BusPRIME defined an BPMN extension that provided three major improvements from the business perspective.

1. It allowed the businesses to annotate the use of data in the processes that were using it. In particular it allowed the use of Personal Identifying Information (PII as defined under GDPR) to be recorded in the business processes. This meant that the business process designers were aware of the use of personal data, and the conditions governing the use of the data, thereby reducing accidental infringements.
2. It allowed the automatic generation of enforcement policies to enforce the restrictions on the use of personal data. For BusPRIME this was achieved through the automatic generation of XACML¹⁶⁰ policies from the business process definitions. This meant the use of data and its access restrictions were tied together, with no possibility of an accidental usage of the data.
3. It allowed the generation of up-to-date GDPR statements. Often the GDPR statement is drawn up from "second hand" or incomplete knowledge of how the data is actually used. Should a business change suppliers or partners over time, this may not have been captured accurately in the GDPR statement. Using BusPRIME and the associated business processes GDPR statements can be automatically updated to match the current business processes.

Privacy-by-design is a laudable principle, however in practice it is tricky to implement. BusPRIME provided a means to achieve privacy-by-design in a way that was business friendly and could be automatically enforced.

3.4.2 Polirural: SEMEX

SEMEX (Semantic Explorer)¹⁶¹ is one of the outputs of the Polirural¹⁶² research project. One of the objectives of the Polirural project is to bring solutions to policymakers to support European rural areas in responding to contemporary challenges. The tool has been developed to provide support to researchers and facilitators by reducing the cognitive load related to tasks that are essential to policy processes, such as Foresight and Policy evaluation

¹⁵⁹ Cavoukian, A., Taylor, S., & Abrams, M. E. (2010). Privacy by Design: essential for organizational accountability and strong business practices. *Identity in the Information Society*, 3(2), 405–413.

¹⁶⁰ OASIS XACML-TC. (2014). OASIS XACML TC webpage. http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

¹⁶¹ <https://semex.io>

¹⁶² <https://polirural.eu>

It is an open-source web app tool based on cutting edge text mining (TM) technology, capable of extracting information from unstructured data and displaying the results through graphical and textual outputs. The tool is based on NLP functions including Topic Extraction, Named Entity Recognition and Sentiment Analysis. Through these NLP processes users can access analysed text and visualise the result through clear graphical representations.

Semex.io is the frontend application of a complex backend system based on text mining that is available through open API. The portal is the gateway to a large library of more than 4000 documents related to European rural areas. Sources have been gathered by Polirural Pilots ensuring a high level of expertise in rural related topics and wide geographical representation (12 Pilots in different parts of Europe plus Israel). The scientific articles, technical reports and policy related documents collected relate to the needs of Polirural areas as well as to the local, regional, national and European policies.

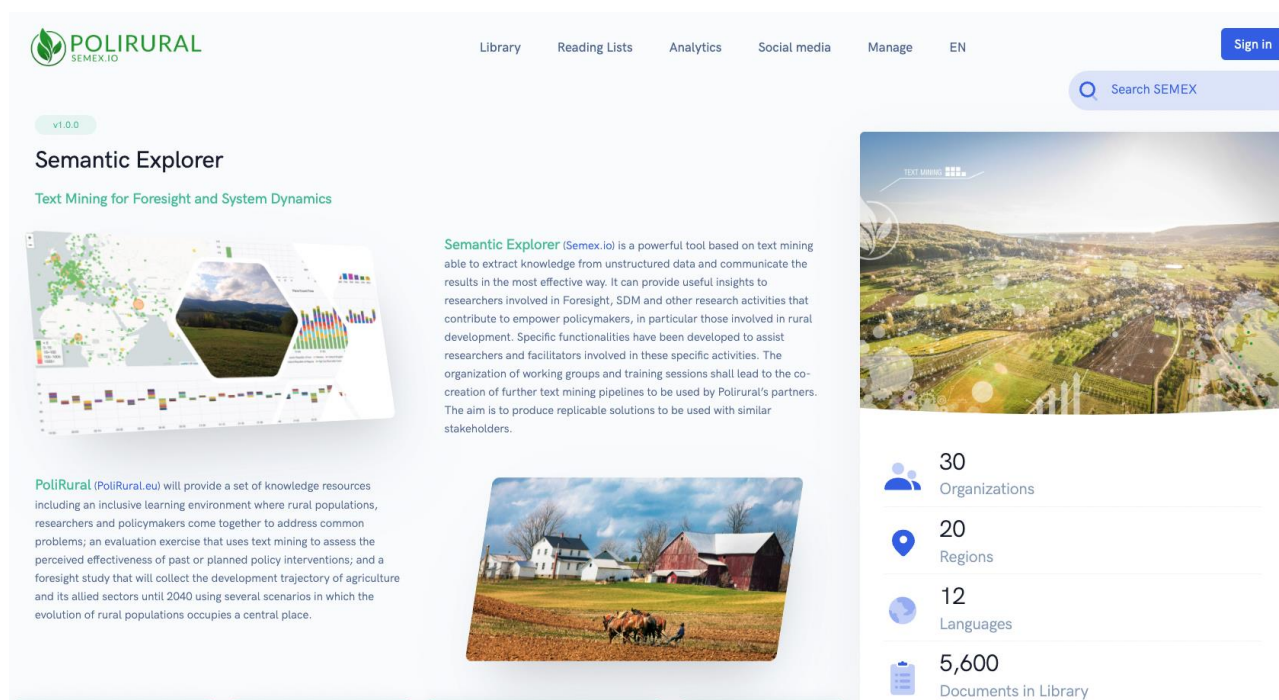


Figure 7. SEMEX Portal (semex.io)

SEMEX can provide useful insights to researchers involved in Foresight, SDM and other research activities that contribute to empower policymakers, in particular those involved in rural development. Specific functionalities have been developed to assist researchers and facilitators involved in these specific activities. The organisation of working groups and training sessions leads to the co-creation of further text mining pipelines to be used by Polirural's partners. The aim is to produce replicable solutions to be used with similar stakeholders.

Natural Language Processing (NLP) breaks down language into shorter, more basic pieces, called tokens (words, periods, etc.), and attempts to understand the relationships between the tokens. In semex.io NLP is used for Keyword and Topic extraction and modelling, Sentiment analysis and Named Entity Recognition (NER). All these tasks require more basic NLP tasks such as Word and Sentence tokenization, Dependency parsing, Part of Speech tagging, lemmatization, semantic comparisons of words.

Semex.io uses a predefined thesaurus called **GEMET (GEneral Multilingual Environmental Thesaurus)**. This thesaurus was developed by the European Environment Agency (EEA) and the European Topic Centre on Catalogue of Data Sources (ETC/CDS) as an indexing, retrieval, and control tool. It contains more than 5000 topics in 37 languages. Topics have hierarchical relationships (broader and narrower terms) and "related term" relationships. In Semex, we choose the topics from GEMET because it was conceived as a "general"

Project 'cities2030' | H2020 ID | 101000640 | 'Co-creating resilient and sustainable food systems towards FOOD2030' | www.cities2030.eu

thesaurus, aimed to define a common general language, a core of general terminology for the environmental sector, which is closely related to our main topic of interest, rural areas. The image here below, extracted from the 'Topic Explorer' functionality, represents topics and subtopics from GEMET related to 'common agricultural policy'.

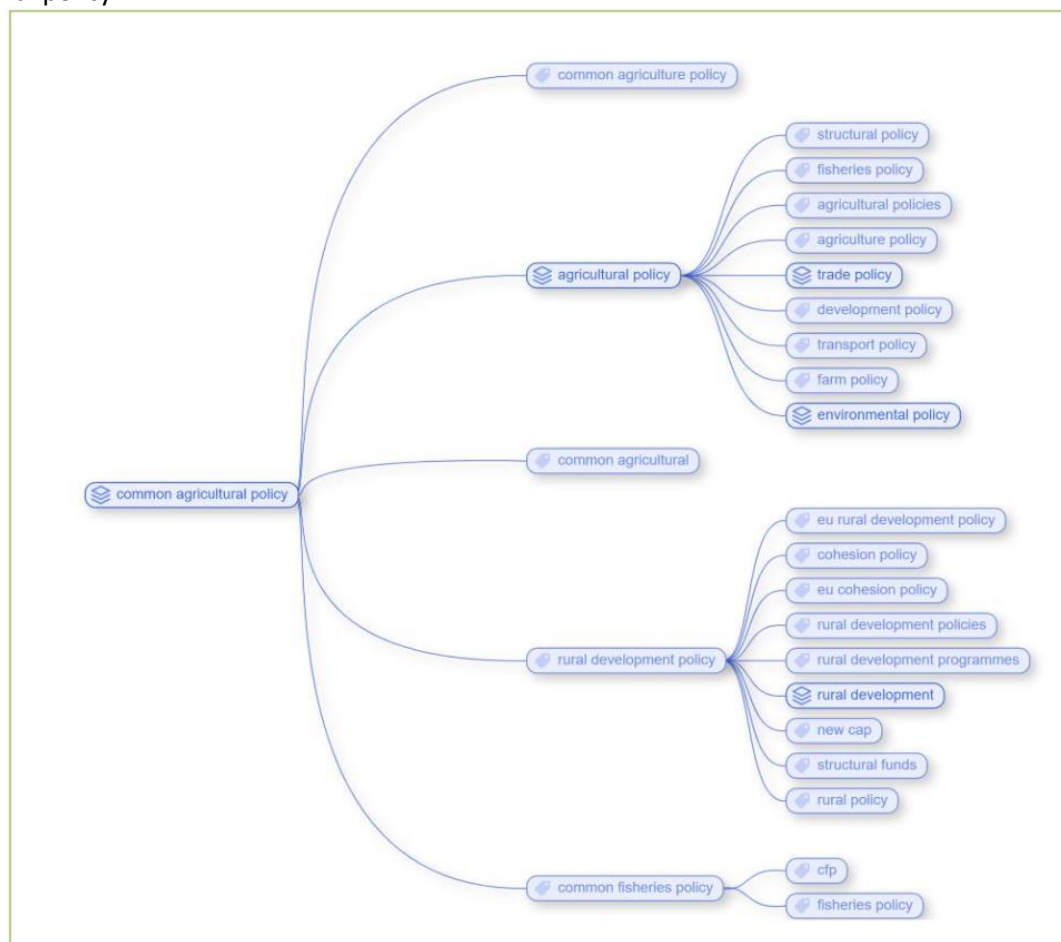


Figure 8. GEneral Multilingual Environmental Thesaurus

The core of the NLP system consists of four main processes:

- **Topic extraction and modelling:** This is the most complex NLP task in the system and is based on multi-label classification. The system uses a restricted list of topics which can be assigned to the text.
- **Keyword extraction:** The most important words or noun phrases are identified in the text. Graph-based TextRank algorithm (based on Google's PageRank algorithm) was selected as the best method to extract a list of keywords from the text. These keywords are also used to produce the summary of each source and of Curated Reading Lists.
- **Topic similarity comparison:** Extracted keywords are compared to a set list of topics in GEMET and the most similar topics are selected. The comparison is done by Word Mover's Distance1 (WMD) algorithm which is one of the most accurate algorithms available to semantically compare documents.
- **Named Entity Recognition (NER):** It is the identification of real-world entities such as persons, organisations, locations and others (with 19 different categories of entities). Extracted entities can link the text to specific geographical location, to an organisation or concrete persons.
- **Sentiment analysis (opinion, emotion):** It is a technique used to identify or classify the polarity of text. The opinion polarity can range from negative (-1), through neutral (0) to positive (1). The

methods used for sentiment analysis can be based on a dictionary of polarity words (terms such as 'waste' are considered negative) or training a machine learning or deep learning model. The Semantic Explorer can process texts in the 10 languages. It was not possible to cover Hebrew and Macedonian because of no availability of trained models and lack of linguists in the project.

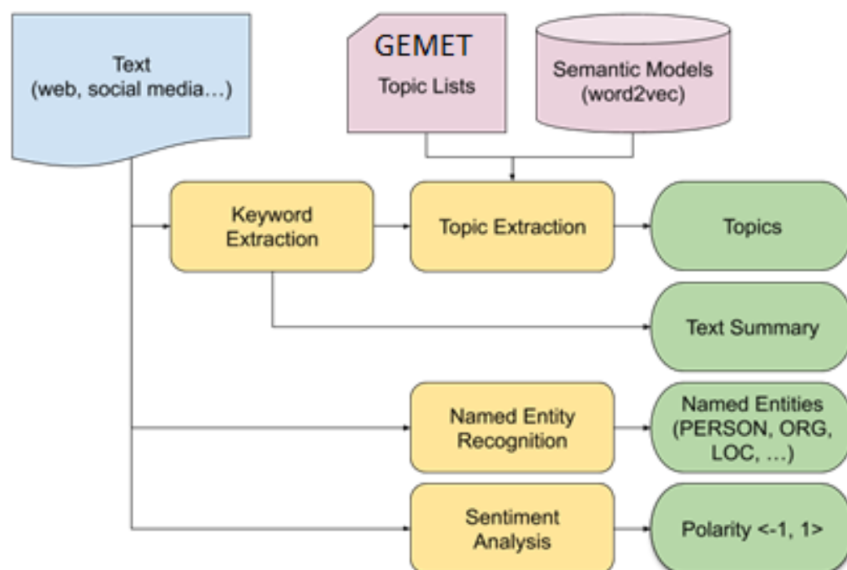


Figure 9. SEMEX NLP core system

A detailed user guide about SEMEX can be found [here](#).

3.4.3 DEMETER

DEMETER¹⁶³ is a project funded by the European Commission which is focused on the digital transformation of the agri-food sector for the benefit of farmers and the industry as a whole.

DEMETER's ambition is to facilitate and speed-up the deployment of data-driven, smart farming solutions that provide decision support and control systems for the agricultural sector. This empowers farmers to make better decisions, allowing them to harness the full value of their own data and knowledge as well as those shared with others. DEMETER aims to do this by enabling data sharing and use between different services, technologies and providers, providing farmers with solutions that are applicable to how they farm regardless of production systems, farm size or location.

DEMETER proposes an overarching approach that integrates heterogeneous technologies, platforms, and systems, while supporting fluid data exchange across the entire agri-food chain, addressing scalability and governance of ownership. Its architecture consists of a DEMETER Provider and a DEMETER Consumer service based on the architecture model introduced by the Industrial Data Space (IDS)¹⁶⁴, then further specified by the International Data Space Association (IDSA)¹⁶⁵, which is the continuation of IDS. As data interoperability is such high importance, the proposed solution provides the necessary data translation mechanisms combining the use of a semantic data model called Agriculture Information Model (AIM), developed by

¹⁶³ <https://h2020-demeter.eu/>

¹⁶⁴ <http://www.industrialdataspace.org/>

¹⁶⁵ <https://www.internationaldataspaces.org/>

DEMETER, along with the respective data translation, management and inference mechanisms adopting widespread standardised solutions (e.g., NGSI-LD¹⁶⁶, Saref4Agri¹⁶⁷, ADAPT¹⁶⁸).

3.4.3.1 Scope of the data-sources used in DEMETER pilots

DEMETER has a total of 20 pilot (open-call) projects underway, which use a variety of data gathered from multiple and somewhat varied sources.

Data gathered includes:

- **Geospatial data.** Date/time, location, (European Global Navigation Satellite System (EGNSS), the European Geostationary Navigation Overlay Service (EGNOS), the European global satellite-based navigation system, etc.), Earth Observation, thermal imagery, multispectral imagery derived by drones – Unmanned Aerial Vehicles - UAV, PlanetScope hi-res, satellite data (Copernicus), LPIS, 3D point clouds, etc.
- **Machinery data.** Exhaust temperature, NOx-Conversion, machine telemetry, engine data and after treatment data.
- **Farm/Logistics data.** Historical farm statistics, FADN, water consumption, energy consumption, fuel consumption and amount of fertiliser.
- **Environmental data.** Air temperature, air humidity, wind speed, wind direction, solar radiation and precipitation.
- **Water data.** Water salinity, water temperature and water height.
- **Soil data.** Soil temperature, soil salinity, humidity, soil conductivity, soil water tension, soil water content and nutrients monitoring/control data.
- **Crop data.** Planting details (e.g. date), HyperSpectral signature, irrigation data, water pumping data, the Normalised Difference Vegetation Index (NDVI), pest info, and others.
- **Animal welfare data.** Beehives data, animal health data, animal respiration data, eating habits data, lameness detection and grazing time.
- **Product data.** Product quality data, milk quality data, milk composition data and milk yield.

3.4.3.2 Common Data Models and Semantic Interoperability Mechanisms

DEMETER does not define completely new interoperability mechanisms but instead uses (and extends) a wide range of pre-existing mechanisms at sensor, data, and service levels. The DEMETER Data View primarily is supported by two architectural components (referred to as

Core Enablers):

- **Data & Knowledge Repository.** This component targets the problem of the physical representation of the semantic model. It creates a mapping between different communication, networking, and digital resources which generates the information flow between the business processes.
- **User & Resource Registry.** This component represents a relational database management system that structures data in tables with certain properties. The database will contain information relating to users, credentials, access controls and views etc.
- **Semantic Interoperability Enabler.** This core component represents the "translation" knowledge for automating the translation of "raw" data from other FMIS / IoT / Machinery platforms and data sources to an AIM compliant model.

¹⁶⁶ ETSI. Context Information Management (CIM); NGSI-LD API. Available online: https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.01.01_60/gs_CIM009v010101p.pdf

¹⁶⁷ ETSI. SmartM2M; Extension to SAREF; Part 6: Smart Agriculture and Food Chain Domain. Available online: https://www.etsi.org/deliver/etsi_ts/103400_103499/10341006/01.01.01_60/ts_10341006v010101p.pdf

¹⁶⁸ ADAPT - Agricultural Data Application Programming Toolkit, Homepage URL: <https://adaptframework.org/>

A keystone of the DEMETER approach is to combine the use of a semantic data model (the DEMETER Agriculture Information Mode — AIM) along with the respective data translation / management / inference mechanisms adopting widespread standardised solutions. This model serves two purposes. First, it makes data from different sources available through the interoperability of existing databases. Second, it provides access to new data that can address evaluation questions for significant policy interventions despite not being always directly linked.

DEMETER's Agricultural Information Mode¹⁶⁹ defines a set of mappings between DEMETER's own ontology, and FIWARE ; Saref4Agri ; ADAPT ; INSPIRE and FOODIE; AGROVOC ; EPPO ; and Earth Observation standards.

3.4.3.3 Analysis of the approach

While the approach taken in DEMETER is valid and they have demonstrated its applicability in a series of pilots, it suffers from the complexity of using a meta-model to perform the translation. This adds a layer of indirection to the conversion process, which adds meta-data parsing and processing time to the data conversions themselves. Having such a flexible mechanism that can translate any type of data type to any other data type is overkill for most use cases.

The argument being that once a data-source is identified it is only capable of producing certain types of data. The requirement being to translate from one very specific data format to a specific target data format, both of which are known beforehand. Other sources of challenges when attempting a translation include:

1. A data-source may not comply fully with the equivalent standard (ie. be slightly non-compliant)
2. A data-source may not support all the concepts or attributes of a standard (ie. compliant but supports a subset of standard)
3. A data-source may not have a standard to comply with. This means it is a de-facto standard in its own right.
4. Translating all possible attributes of one data format to another, is inefficient if only a handful of attributes are required.

A second argument is the amount of effort to produce these standardised mappings. Where the number of input standards (for data formats) and the number of output standards (for data formats) are small, mappings are possible. However, this does not scale up, as the amount of effort required increases the more standards that are supported. This means a (n X n) set of translations required, where n is the number of data-formats supported, which for large values of n is unsustainable.

For CITIES 2030, this mapping can be used for some data-sources, however CITIES2030 uses a much larger variety of data-sources than those defined in DEMETER. Based on interaction with some of Living Labs, in most cases the data recorded is non-standard in the sense it has a defined format and representation but may not comply with any specified standard (due to the absence of an agreed standard). This implies that this approach is useful but limited in real applicability given the number of data-sources that need to be addressed by CITIES2030.

¹⁶⁹ Demeter, "D2.3 Common Data Models and Semantic Interoperability Mechanisms." DEMETER H2020, 2020. <https://h2020-demeter.eu/deliverables/>

4 Data integration and analysis subsystem: architecture and design

Data integration and analysis is referred to as "Data-science", and is sometimes defined as "data science is the science of data" or "data science is the study of data"¹⁷⁰. Others define it as "the field of applying advanced analytics methods and scientific concepts to derive useful business information from data"¹⁷¹. Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

In this section we described the S2CP platform's subsystem focused on providing such functionalities and services. First, we introduce the general functionalities of the data integration and analysis subsystem, and (secondly) we describe the specific components in the S2CP platform that cover those functionalities.

4.1 Introduction to data integration and analysis

Data-science fundamentally, is a support activity to ask questions and see what data is available to support answers to those questions. The process of doing data-science can be visualised as follows.

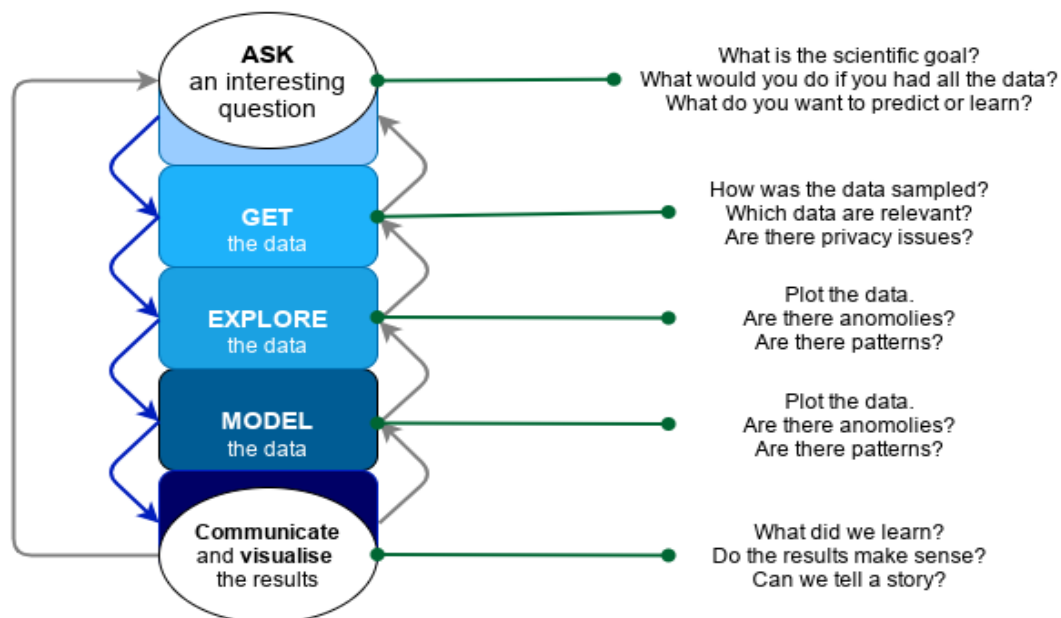


Figure 10. Data science process: Adapted from: Blitzsten, J. and Pfister, H., Harvard data science course CS109.

The data used for analysis can come from many different sources and presented in various formats. This stage is sometimes referred to as data-discovery. There are also data-governance issues with the processing, handling and storage of certain data.

Once all the data-sources have been identified, then the data can be explored. Key tasks undertaken include the identification of anomalies (which may skew results) and the identification of features that are of interest.

¹⁷⁰ Cao, L. (2017). Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1–42.

¹⁷¹ Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science, 2(5), 1–22.

Data science uses complex machine learning algorithms to build predictive or classification models. While basic analytics offer a description of data in general, more advanced analytics can offer a deeper understanding of data by using granular data. The granular examination of the data can provide insights which may answer the question posed.

4.1.1 A generic data processing solution

Taking a process viewpoint the Data Integration and Management sub-system should cover the following steps:

- **Data gathering:** Where the data is aggregated from the various sources to a convenient location to enable future processing.
- **Data management:** Here the data is actively managed to ensure any and all restrictions or obligations as to the usage of the data are enforced and the data is not tampered with.
- **Data discovery:** Where the existence of data-sets and the information there-in can be discovered and explored. This most often requires combining different data-sources to gain enough information to permit further analysis. The collated data then goes through two modular components in a pipeline to achieve the underlying objective, to provide insights, and business intelligence to different actors. These combined are referred to as data-pipelines and are used to feed the visualisation component.
- **Data processing:** Data processing consists of transformation required on the data before it is passed in the pipeline to analytical models. The methodology here consists of approaches such as data clustering and feature engineering.
- **Data analysis:** The data analytics pipeline consists of approaches such as ensemble learning to deal with the heterogeneity in data to address different questions of all the actors involved.

4.1.1 Pipelines for CRFS

Based on a survey of the Use-cases defined and the necessity of providing a "Single Click" solution for answering these questions, a move was made to more specialised data-pipelines each targeting a specific data-source type (eg. Social media information, Geospatial data). The primary reasons for this decision are outlined below:

1. It was easier to explain the purpose of and benefits of using these components to the Living Labs.
2. It is easier to integrate a pre-baked data-pipeline to answer specific questions on the data. Both the data-processing / cleaning decisions have been made, and an algorithm selected which based on expertise and previous research work, best answers the question posed.
3. Less work is needed to re-use them between use-cases.
4. These pipelines are optimised to handle the specific data needed to provide insights.

For these reasons, pre-baked pipelines focused on specific types of questions and handling specific types of data have been chosen. These are outlined in the following sections.

4.2 Data integration and management subsystem: architecture and components

The data integration and analysis subsystem is one of the core components of S2CP platform that enables different aspects related with data via collection of developed microservice, components and API exposition. S2CP data integration and analysis subsystem consists of the following services and components (in purple), to be used by different CRFS actors.

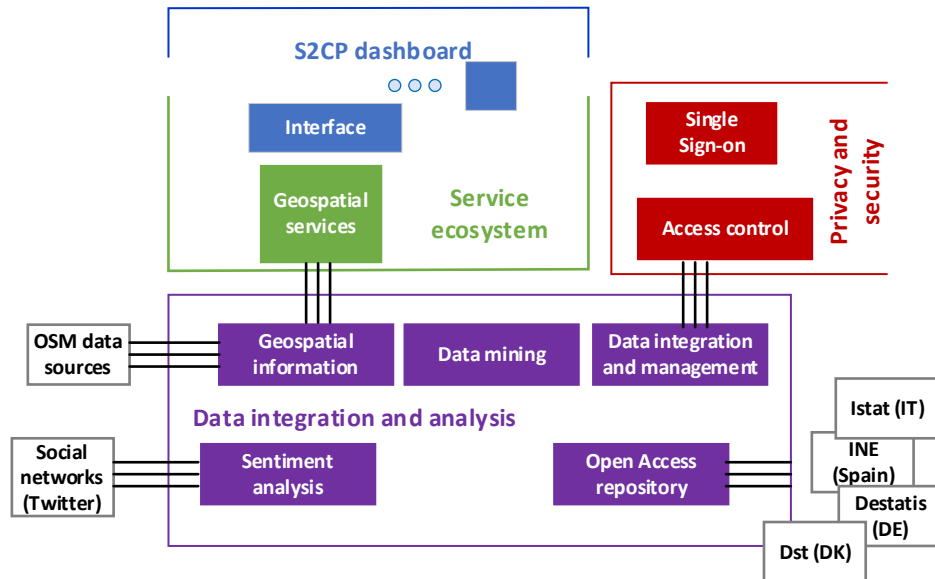


Figure 11. Data integration and analysis subsystem architecture

4.2.1 Integration and management service

The integration and analysis service is responsible for data acquisition from different resources and putting it in CITIES2030 data model format. The data coming for analysis would have different data models, and the **integration** service will create a central database view to put data from all the sources into a data warehouse cloud location.

The collated data then goes through two modular components to achieve the underlying objective and provide insights / business intelligence to different actors. These combined are referred to as data-pipelines and are used to feed the visualisation component.

The management aspect is handled by the Data Governance component which is further defined in D6.4. However, the supporting infrastructure and meta-data to enable data-governance is provided within this service.

This is a summary of data handling steps covered.

Data steps:	gathering	management	discovery	processing	analysis
-------------	------------------	-------------------	------------------	------------	----------

And an infographic representation of this same is provided below.

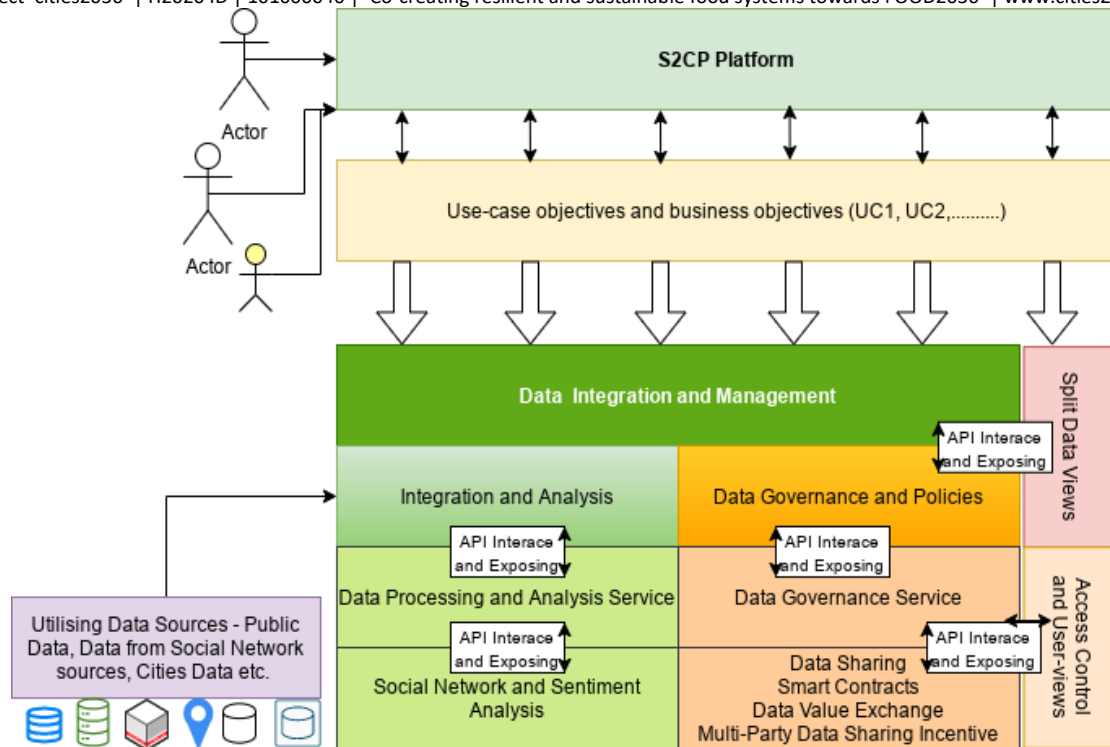


Figure 12. Interactions on the Data Integration and Management component

4.2.2 Geospatial information

This subsection describes the *Geospatial information component*, which collects relevant information from external geospatial sources. This component stores geographical distribution of various CRFS agents in a relational database, considering many attributes that are later explained in Section 6.

The following functionalities have been taken into account for the development of this component:

- Consideration of geographical places at city level, currently supporting 10 cities (CRFS labs).
- Definition of workflow for the update of information. The workflow takes into account currently available information from Open Street Maps public database.
- Filtering by amenities and shops, divided into categories, to obtain a spatial vision of the distribution and relevance of locations or points of interest.
- Possibility of personalised visualisations, thanks to the integration into the Geospatial CRFS Web Services component of T6.3 and the S2CP dashboard (T6.5).

4.2.3 Open access repository

Publishing and consuming open data is a cornerstone for the development of applications and the creation of an innovation ecosystem. In this regard, this section explains how Cities2030 data repository has been created, and how users can expose their data publishing it in this S2CP component.

Cities2030 data repository is an open-source solution based on the well-known Open Data Publication network CKAN¹⁷² (Comprehensive Knowledge Archive Network), most widely used by cities, public authorities, and organizations. The following functionalities are considered:

- Enable the publication, management, and consumption of open data, usually, but not only, through static datasets (CSV, XLSX, etc.).

¹⁷² <https://ckan.org/>

- Possibility to catalog, upload, and manage open datasets and data sources, while supports searching, browsing, visualizing, or accessing open data.
- Relation to the Geospatial Information component, since the data that this component manages to generate will be published in the Open Access data repository by all labs.
- Possibility to plot the data into an interactive map.
- Filter of results by geographical location, specifying a bounding box to limit the interesting areas. Different coordinate geometries and formats are supported.
- Extensible architecture, making it easy to support other standards and distribution services.

4.2.4 Social Network and Sentiment analysis component

Collects data from social networks, for example Twitter, and further performs statistical, textual, and image analysis with the aim to understand the processes and opinions of the users. This component has the following functionalities:

- Collect posts corresponding to certain queries specified by the user.
- Calculate statistics about the occurrence of various terms in posts.
- Perform sentiment analysis of the associated text.
- Perform image detection (identification of objects present on images).
- Summarise the collected information in the form of plots showing historical and geographical distributions.

4.2.5 Data mining tool

The data mining platform allows analyzing trends in society in relation to food systems using information sources such as social networks, personal blogs, or internet forums. The following functionalities are expected from this component:

- Analysis of the typology of a website to detect if the theme is related to food.
- Determination of weights to obtain how much food-related content is on the website, its relevance.
- Possibility of plotting the results over time, to assess the perception trend for each food group.
- Integration of other techniques such as sentiment analysis, which are also part of the data integration and management subsystem.

5 Social network and sentiment analysis component

The aim of this component is to infer the opinion of consumers about various activities and actions taken by the LABs. To achieve this, the component gathers information related to the relevant actors from social network Twitter and performs sentiment analysis of the associated information (text of tweets and associated images), see figure below. The data is collected from the Twitter Universe and further filtered to ensure the relevance to the studied question. Next, the sentiment analysis is performed with the help of the latest technologies available in NLP (Natural Language processing) and Image Processing domains, namely the Artificial Neural Network (ANN) -based algorithms. After that, the results are aggregated and summarised. Finally, they are presented in the form of plots and tables for further analysis by a user. This information can later be used *to understand the opinion of citizens about different topics, track the occurrence of certain projects on social networks, plan and correct action plans etc.*

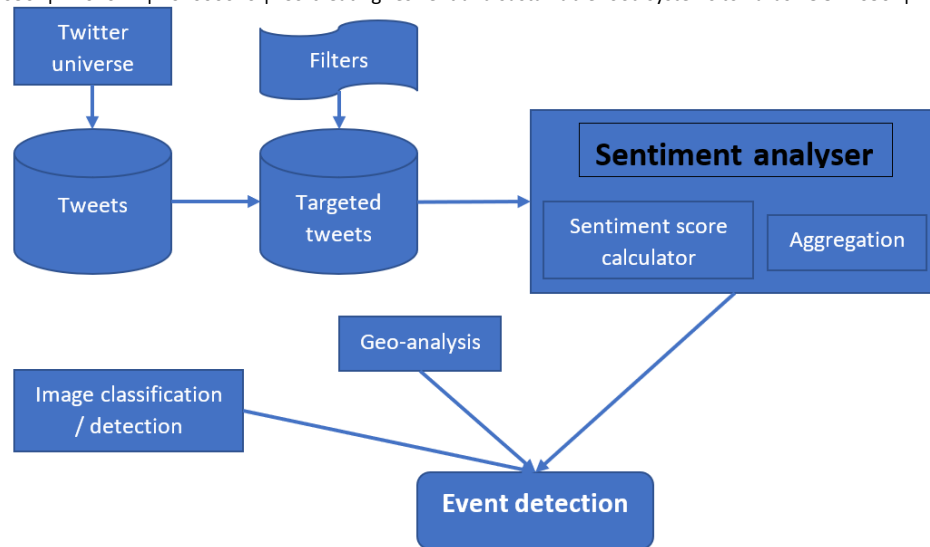
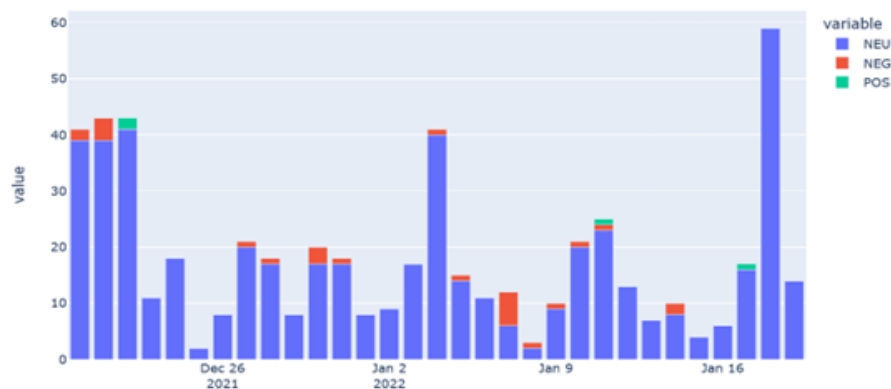
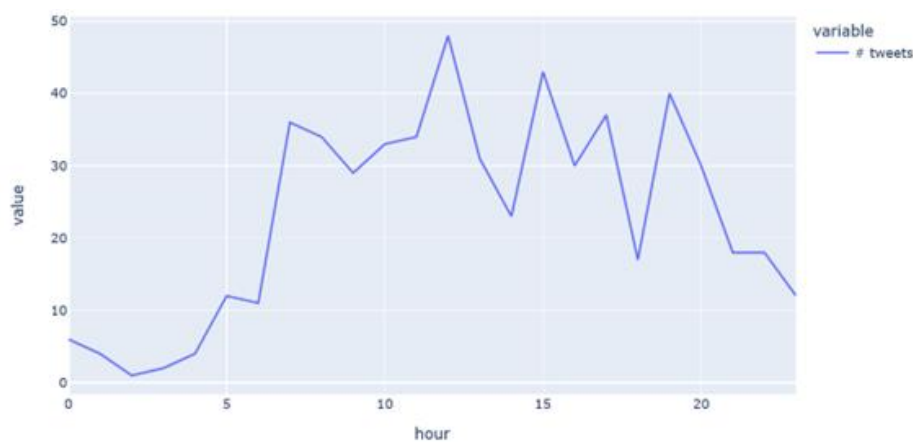


Figure 13. A schematic representation of the component

The provided visualisation of different granularity can include historical and geographical distribution of tweets of different sentiment (Positive, Neutral and Negative).



a) Per-day distribution



b) Per-hour distribution

Figure 14. Historical distribution of tweets neutral (NEU), negative (NEG) and positive (POS) tweets

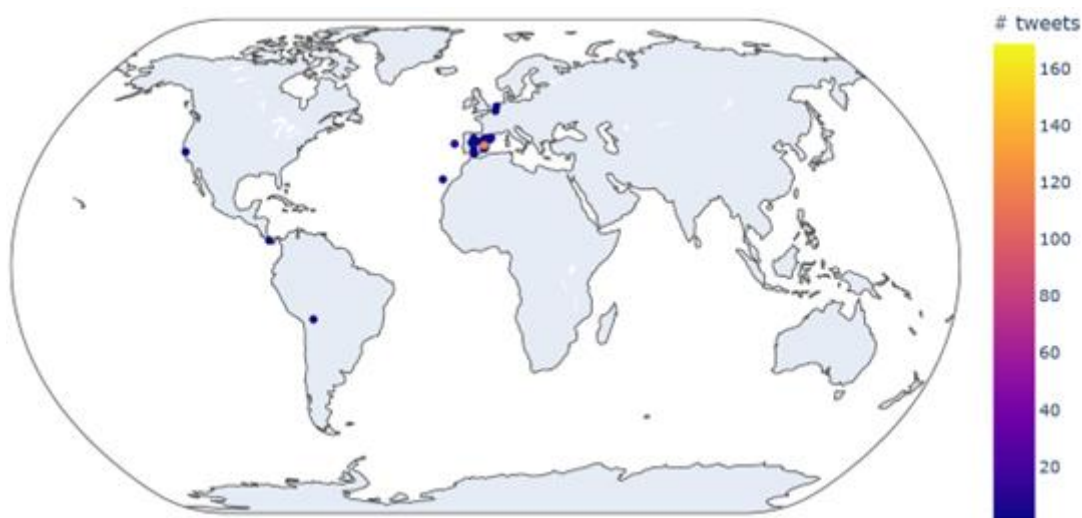


Figure 15. Geographical distribution of tweets

Additionally, the component can perform the analysis of the images associated with the tweets to better target specific events/actions/discussions. A relevant example is shown below.



Top predictions:

grocery_store: 0.80

banana: 0.10

orange: 0.03

Figure 16. Results of image analysis

The code of the component can be found at the following repository¹⁷³.

To ensure the reliability of the generated results, we employ the techniques from conformal learning and causal analysis. As it was described in the [Conformal learning](#) section, this framework is designed to ensure user-defined accuracy requirements of a predictive model. This means that we can choose a threshold, for example 95% accuracy, for the predictive system that is being built. We use this framework at the sentiment prediction and image classification/prediction steps. Our experiments with conformal learning techniques

¹⁷³ <https://github.com/marharyta-aleksandrova/cities-sentiment>

Project 'cities2030' | H2020 ID | 101000640 | 'Co-creating resilient and sustainable food systems towards FOOD2030' | www.cities2030.eu resulted in a study about the impact of nonconformity functions on the efficiency of conformal classifiers. The letter resulted in 2 publications^{174,175}.

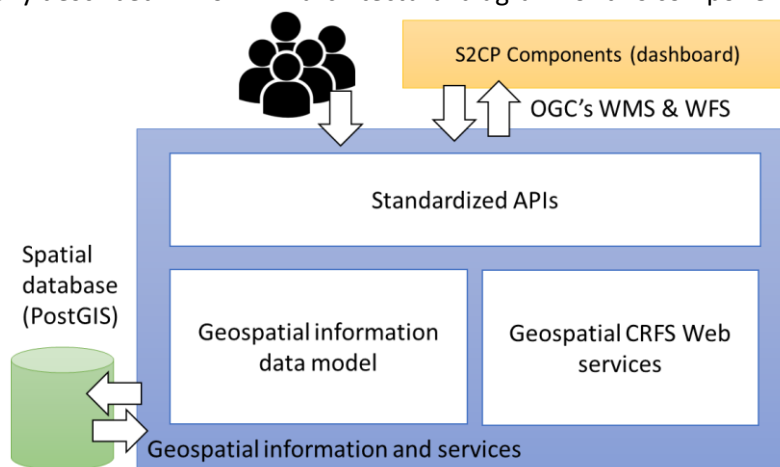
Traditional Machine Learning techniques are good at revealing associations between attributes. However, when data is analysed to make some decisions what is required is causal relationships. The theory of causality aims to fill in this gap. In the sentiment analysis tool, we aim to use the causality theory, see [Causal analysis](#) section, to set up the analytical process to be able to find causal links and measure the strengths of relationships. Our attempts to incorporate the relevant techniques into the developed tools resulted in a study of the noise impact on the efficiency of RESIT approach which was presented at a conference¹⁷⁶ and published as a journal paper¹⁷⁷.

6 Geospatial information

The Geospatial information data model is based on state-of-the-art technologies and common and standardized application interfaces in commercial products and in the world of research. This component develops a data model and a repository containing the geographical distribution of the different agents throughout the different CRFS.

6.1 Architecture and process

The Geospatial information data model is part of the Geospatial information and services component, which is a S2CP entity briefly described in D6.1. An architectural diagram for this component is shown below:



¹⁷⁴ Aleksandrova, Marharyta, and Oleg Chertov. "Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study." *Conformal and Probabilistic Prediction and Applications*. PMLR, 2021.

¹⁷⁵ Aleksandrova, Marharyta, and Oleg Chertov. "How Nonconformity Functions and Difficulty of Datasets Impact the Efficiency of Conformal Classifiers." *arXiv preprint arXiv:2108.05677* (2021). Presented at DFUQ workshop, ICML-2021.

¹⁷⁶ Kap, Benjamin, Marharyta Aleksandrova, and Thomas Engel. "Causal Identification with Additive Noise Models: Quantifying the Effect of Noise." *arXiv preprint arXiv:2110.08087* (2021). Presented at 10èmes Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes (JFRB-2021),

¹⁷⁷ Kap, B., Aleksandrova, M., & Engel, T. (2021, November). The Effect of Noise Level on the Accuracy of Causal Discovery Methods with Additive Noise Models. In *Benelux Conference on Artificial Intelligence* (pp. 120-140). Springer, Cham.

Figure 17. Architecture of Geospatial information and services component

This entity collects relevant information from external geospatial sources and stores geographical distribution of various CRFS agents in a relational database. The following functionalities have been taken into account for the development of this component:

- Consideration of geographical places at city level, currently supporting 10 cities (CRFS labs).
- Definition of workflow for the update of information. The workflow takes into account currently available information from Open Street Maps public database.
- Filtering by amenities and shops, divided into categories, to obtain a spatial vision of the distribution and relevance of locations or points of interest.
- Possibility of personalized visualizations, thanks to the integration into the Geospatial CRFS Web Services of T6.3 and the S2CP dashboard of T6.5.

For each city a couple of datasets are supported according to OpenStreetMaps formats¹⁷⁸:

- Point-type set (OSM nodes), which represents the point-based elements of the establishments according to some categories.
- Linestring type set (OSM Ways), which are presented in open or closed polygons, generally demarcating an area of an establishment, such as a supermarket.

The Geospatial information and services component describes a workflow that must be followed to obtain spatial information, which ranges from the collection of information in public repositories to user engagement in the use and update of information. Below is a diagram of this process:

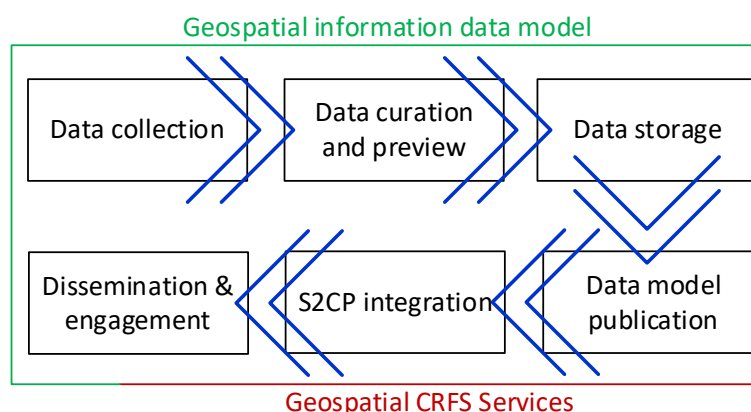


Figure 18. Geospatial information and services component workflow

In the *first phase* we find the data collection process. This process will be in charge of designing the necessary mechanisms for the creation of tools that allow automating the downloading of information of interest, in relation to each city. It also defines the data update mechanisms that must be followed to finally keep information always up to date in the repository. This process will be described in Section 6.2.

In the *second phase* we find the data curation and preview. This process consists of accessing the information downloaded in the previous step, and validating that it meets the conditions that ensure its spatial coherence. This process will be described in Section 6.3.

In the *third phase* we find the Data storage function. In Section 6.4. the tools used for this storage and the data model used are described.

The *fourth and fifth phases* (Publication and S2CP integration) belong to the Geospatial Web Services component, and therefore will be described in more depth in the final report of T6.3, since it is more related to this task than to T6. two.

¹⁷⁸ https://wiki.openstreetmap.org/wiki/OSM_file_formats

The *sixth phase* is related to dissemination and user engagement, where the activities proposed to make this component known and to achieve acceptance and use by the CRFS labs will be described. This phase will be detailed in Section 6.5.

6.2 Data collection

To download points of interest from OpenStreetMaps, a python script has been created. It has been necessary to previously install the following python libraries for data integration and analysis: *overpass*, *geojson*, *requests* and *osmtogeojson*. When selecting the required data, a series of labels are used, such as amenity and shop. The following figure is a code snippet for the process to obtain point-type sets for the city of Bremerhaven.

```
api = overpass.API(timeout=500)
res = api.get("""
area['name'='Bremerhaven']->.city;
(
  node["amenity"="bar"](area.city);
  node["amenity"="pub"](area.city);
  node["amenity"="bbq"](area.city);
  node["amenity"="biergarten"](area.city);
  node["amenity"="cafe"](area.city);
  node["amenity"="restaurant"](area.city);
  node["amenity"="drinking water"](area.city);
  node["amenity"="fast food"](area.city);
  node["amenity"="food court"](area.city);
  node["amenity"="ice cream"](area.city);
  node["shop"="supermarket"](area.city);
  node["shop"="bakery"](area.city);
  node["shop"="butcher"](area.city);
  node["shop"="newsagent"](area.city);
  node["shop"="beverages"](area.city);
);
out geom;
""", verbosity='geom')
```

Figure 19. Code snippet for data collection in Bremerhaven lab

6.3 Data curation and preview

Using spatial operations within the tool we can filter all the points that are not within this territorial zone. Once the data curation is done, we move on to the data preview phase. In this phase it is important to carry out a visual check of the data obtained to verify that it is subject to the following rules:

- The typologies found are correctly differentiated from each other. That is, it is possible to identify immediately the type of point of interest and the regions where points of the same type predominate.
- The chosen icons do not cause the data to be concentrated too much in specific areas. If this is the case, other tools managed in the data integration and analysis subsystem can be used, such as the provision of spatial representation based on clusters.

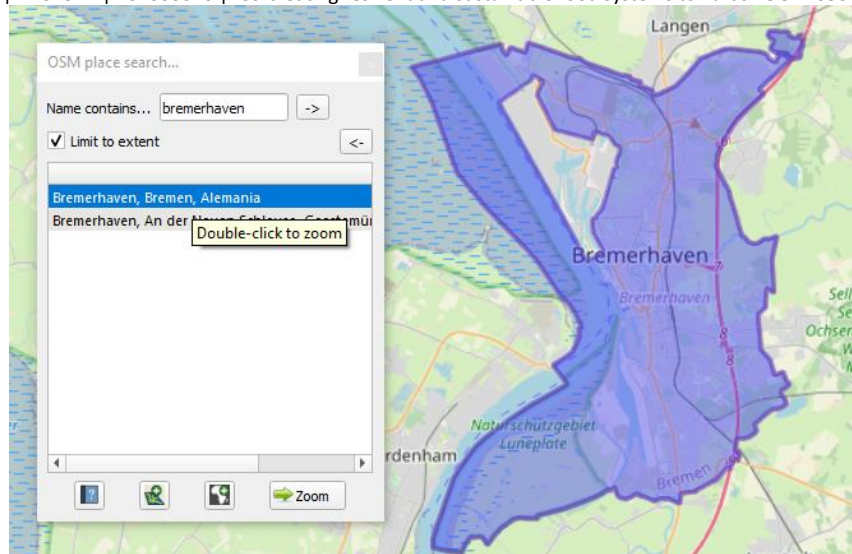


Figure 20. Process for obtaining territorial limit for Bremerhaven lab

To achieve these objectives, an information representation based on the categorization of symbols is proposed, using .svg icons to represent shops and amenities.

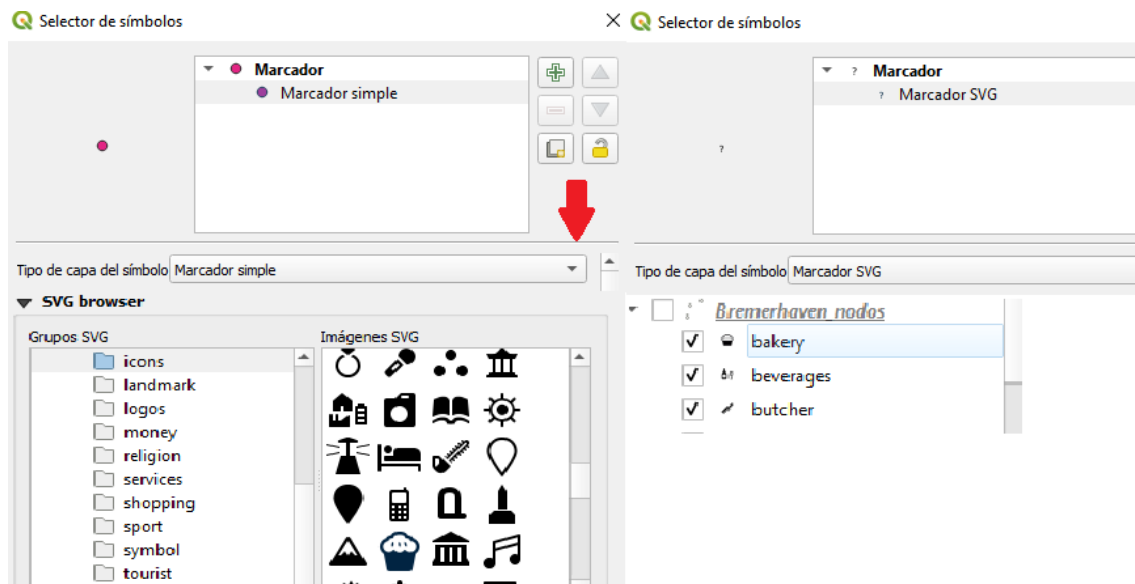


Figure 21. Symbol categorization process in QGIS

The preview result for the city of Bremerhaven is as follows:

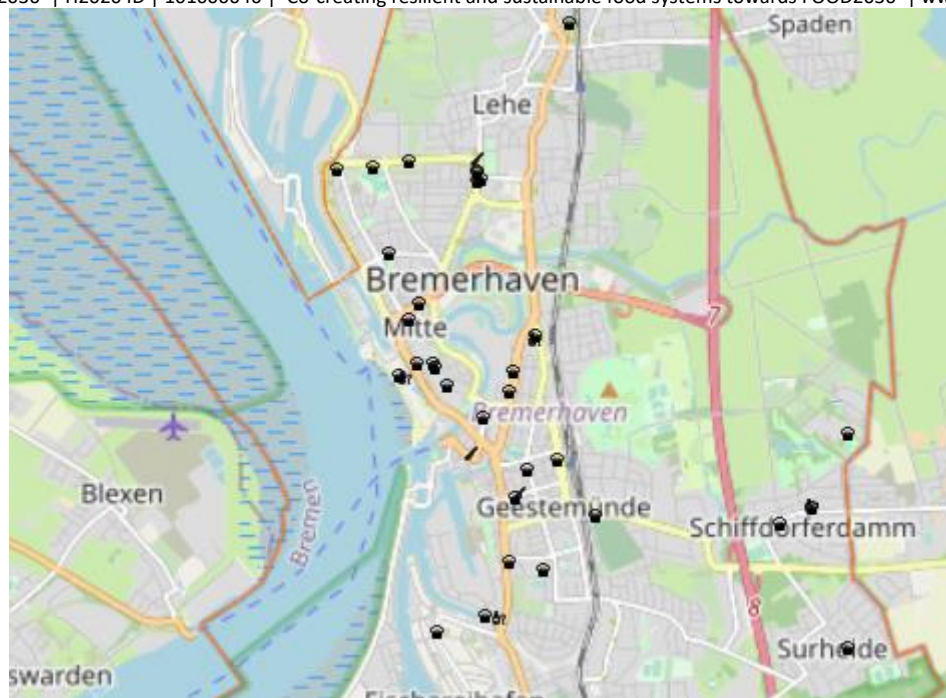


Figure 22. Result of data curation and preview process

6.4 Data storage

For the storage of information and to allow quick consultation without losing versatility, the use of a relational database with spatial extension (PostgreSQL v14 with PostGIS extension) is chosen.

In this phase, the data model to be stored is detailed, where all entities have some **mandatory information**, which is the <id> and the geographic coordinates of the place, as follows:

- Id: Identifier in the database
- Geometry-type: Point | Linestring
- Coordinates: Coordinates expressed in WSG84 standard.

Other properties are optional and are represented in the table of section 7.3.2, describing Data integration and management service data sources.

The result of this phase is a database that represents all the planned geospatial information, following a table structure, such as the following figure, which shows data on amenities in the city of Seinäjoki:

Deliverable D6.2



Query Query History Scratch Pad x

```
1 SELECT * FROM public."Seinäjoki_nodes"
2 ORDER BY id ASC
```

Data output Messages Notifications

	id integer varying	addr.housenumber character varying	addr.postcode character varying	addr.street character varying	addr.unit character varying	amenity character varying
94		57	60320	Suupohjantie	[null]	cafe
95		6	60100	Tehtaantie	[null]	restaurant
96		4	60510	Kaarretie	[null]	restaurant
97		12	60100	Yrittäjätie	[null]	cafe
98		21	60120	Rengastie	[null]	restaurant
99		15	60220	Alaseinäjöentie	[null]	restaurant
100		14	60120	Väinölänkatu	[null]	cafe
101		32	60120	Päivöläkatu	[null]	cafe
102		2	60510	Hyllykalliontie	[null]	restaurant
103		11	60100	Vaasantie	[null]	restaurant
104		25	60100	Nyykoolinkatu	[null]	restaurant

Figure 23. Result of data storage process for Seinäjoki lab

6.5 Dissemination and engagement

It is intended to support all the labs that are interested in this component. See Table 4 of D6.3. To do this, it is intended to verify that all the information pertaining to those cities is available. Once this is verified, we need to download and merge that information about the administrative limits of these cities.

Finally, we will disseminate the operation of this component through a user manual, and a set of events where we will make this tool available to the labs, through the S2CP dashboard.

The workshops where this showcase is planned are the R-213 "Second Workshop with pilots" of the M33, where a preliminary version will be presented, and the R-214 "Third Workshop with Pilots", where a version integrated in the S2CP dashboard will be shown. For more information on these Workshops, please see Gantt chart of D6.1 Section 2.3. The participation of most of the labs or at least of the labs that showed interest in this component is expected.

7 Data integration and management service

A micro-service based design will be used for building this component. The current envisioned micro-service are outlined in the sections below.

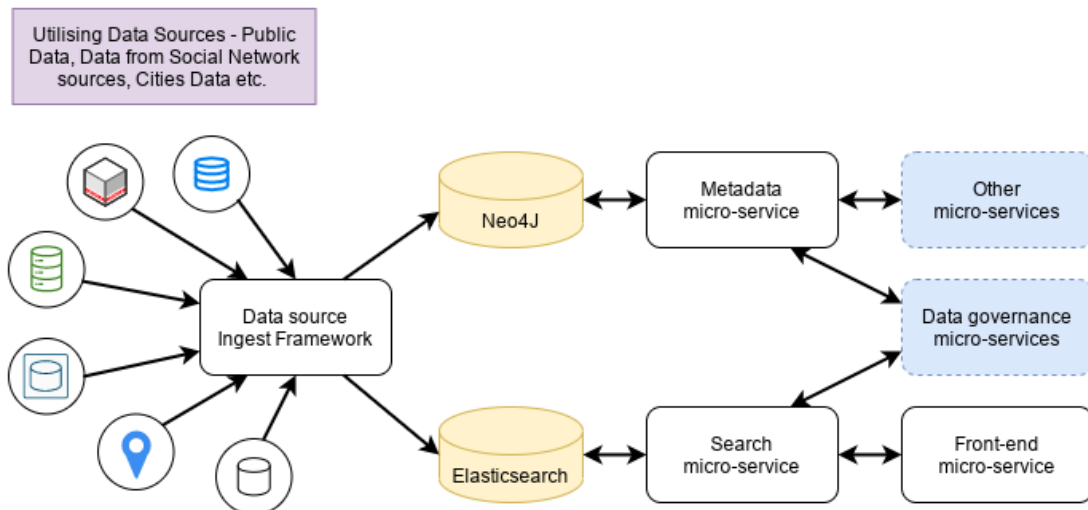


Figure 24. Data Integration, Discovery and Access Micro-services

7.1 Micro-service description

In this first subsection, the most relevant microservices in the data integration component are described.

7.1.1 Front-end micro-service

The micro-service offers a frontend UI for users which will be web based. This will allow end-users, data-scientists to discover available data-sources, query for different types of data available and through the use of the Data Governance component, arrange data-sharing agreements.

7.1.1.1 Implementation

A standard well-known web framework is used, along with well-tested (and known) web frameworks. This allows a larger talent of developers/web designers to customise or rebrand the "look" to suit national, or regional or city preferences. The Frontend micro-service will be a Flask- based web app where the UI layer is built with a standard front-end web stack of React with Redux, Bootstrap, Webpack, and Babel.

7.1.2 Search micro-service

The search micro-service allows data source discovery, by servicing search requests from the Frontend micro-service. It allows searching based on the data-source / table descriptions, as well as by tagged keywords.

7.1.2.1 Implementation

The search index is built with the well-known Elasticsearch's indexer and search functionality. A RESTful API is provided to service search requests from the frontend micro-service. Currently only table resources are indexed and searchable.

7.1.3 Metadata micro-service

This micro-service captures the essence of where particular data comes from, where (in the warehouse of data) it is stored, who is using it, and additional descriptions of its format and purpose. As expected the meta-

data collected matches the meta-data collected for data sources, plus some automatically generated meta-data connections. The metadata is represented as a graph model.

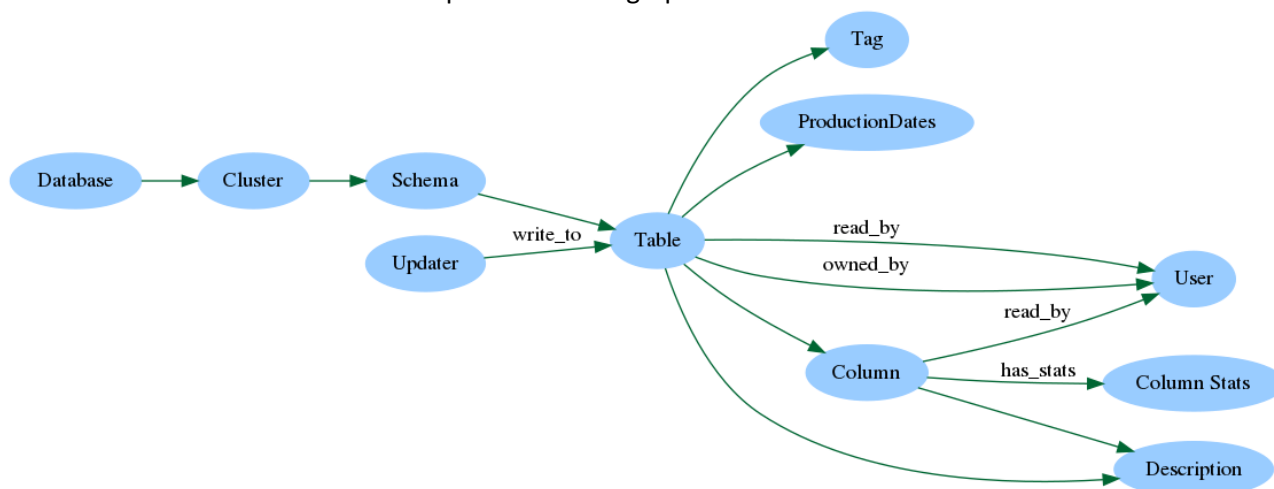


Figure 25. Meta-data structure and inter-relationships

It is important to mention that not all meta-data can be collected for every data-source. For example, production period can only be applied to data that has some time fields present in the data, and hence the corresponding warehouse table.

The meta-data model also allows "tagging" to be applied to facilitate more relevant data-discovery (to support Search micro-service). In addition, the column descriptions will allow the marking of PII within the dataset and corresponding database column.

The primary users of the meta-data micro-service will be the Data Governance component as documented in D6.4 and the Frontend micro-service.

7.1.3.1 Implementation

As the meta-data is inherently Graph based, we will use a Graph DB to represent it. The metadata service will use a Neo4j proxy to interact with Neo4j Graph DB (Java) and serves meta data to support the Search micro-service's search functionality.

7.1.4 Data view micro-service

This micro-service provides access to the raw data contained in the warehouse. In line with the principles outlined for the Data Governance component, data can only be shared or made accessible once a data-sharing agreement is in place. Should a query violate the corresponding agreement, an access privilege error will be generated, and the query will return no results.

7.1.4.1 Implementation

This micro-service will be implemented through a combination of two techniques. First, retrieve the corresponding data-sharing agreement from the Data Governance component. This will be cached to make subsequent queries faster. The second is to re-use of SQL query rewriting capabilities, which are normally used to optimise SQL queries prior to actually executing them. By extending this functionality, table and column names will be checked against the data-view presented data-sharing agreement.

While this may appear novel to Cities2030, Stonebraker & Wong¹⁷⁹ presented this approach to implement access control for databases. Barker & Rosenthal¹⁸⁰ extended the approach to role-based access control by first constructing a view that encodes the access control policy, then rewriting input queries to add WHERE clauses that query the view. Byun & Li¹⁸¹ use a similar approach to enforce purpose-based access control, where purposes are attached to data in the database, then queries are modified to enforce purpose restrictions drawn from a policy. Johnson, Near, Hellerstein, & Song¹⁸² extended this approach further by combining it with noise generating functionality as a means of delivering differential privacy for Uber.

7.1.5 Datasource ingest micro-service

This micro-service focuses on taking the data-source definitions and using to ingest the corresponding meta-data from the data-source. Each data-source is equated to a database table, so this takes the data-source definition and generates a corresponding database schema to support data ingestion. In addition, once data is uploaded to the data-warehouse, it will trigger the collection of automated meta-data extraction, where possible.

This allows a more convenient ingestion of data into the data-warehouse (on-boarding) and automates the updating of the associated meta-data associated with that data.

7.1.5.1 Implementation

This will be implemented as a set of utilities that interact with the micro-service to speed up data-source ingestion. Some of these tools are generic, however, special data may require the use of customised scripts for each data-source. To facilitate maximum flexibility in this regard, it will be implemented as a library of tools for a Linux operating environment.

One example of usefulness of this library, the Data Governance component in D6.4, is expected to re-use this library to try and automate data-governance checks. For example, look for data-source field names containing the work "name". If the corresponding definition does not contain a PII statement, then generate a warning so that the data-source definition is rechecked with the data owner. While this is rather basic checking, it still provides an additional layer of validation of the data-source.

More details will be added once the API is fully implemented.

7.2 Informational architecture

A key aspect of this component is the standardisation of the Data-sources and the meta-information that is expected to come with each data-set. A data-source is a source of specific data for a use-case which will be updated on a specified schedule.

The operation of the data-governance component relies on having data owners for each data-source. Data owners are the ultimately the receivers of any value-exchange¹⁸³ for use of the data. For the purposes of enabling data-governance, the following meta-information is uploaded/collected for each data-source set.

¹⁷⁹ Stonebraker, M., & Wong, E. (1974, January). Access control in a relational data base management system by query modification. In Proceedings of the 1974 annual conference-Volume 1 (pp. 180-186).

¹⁸⁰ Barker, S., & Rosenthal, A. (2002). Flexible security policies in SQL. In Database and Application Security XV (pp. 167-180). Springer, Boston, MA.

¹⁸¹ Byun, J. W., & Li, N. (2008). Purpose based access control for privacy protection in relational database systems. The VLDB Journal, 17(4), 603-619.

¹⁸² Johnson, N., Near, J. P., Hellerstein, J. M., & Song, D. (2018). Chorus: Differential privacy via query rewriting. arXiv preprint arXiv:1809.07750, 30.

¹⁸³ The value exchange term is used as the incentivisation may not include a monetary award.

7.1.1 Meta-data from the Data producer

Title of the data-source: This information includes a clear description of *_what_* the data set covers in a concise fashion. For example, "Milk sales in Bremerhaven".

Business sector: This field captures the business sector covered by the data. It is intended that this field is selected from a limited vocabulary. For the Bremerhaven example, this could be "Food sales".

Geographic-area covered by the data: This field captures the location of the collection of the data. It is intended as a summary field for the geo-distribution of the data contained in the data-set. For most purposes it will be assumed to be a city/town name or city/town hinterland or country region name.

Keywords: This contains a number of pertinent keywords that describe this data. It is intended to augment the information contained in the "Business sector" and title fields. For example, { "Milk", "EDEKA" }

Production period: This will be represented by two date fields (from, to) which cover the range of dates within the data-set. This range can span weeks, months or years. It is recommended that this information be encoded in ISO8601 format eg. (2022-03-01, 2021-05-30) to prevent ambiguities on the date range.

Sensitivity: This field contains a statement on the sensitivity of the data contained in the data-set. It is intended to allow this from a restricted vocabulary. For example:

- Anonymised - Anonymised data derived from personal data.
- Personal - Data set contains personal data, so additional restrictions will apply to its use. ie. Provisions of the GDPR will apply.
- No personal data - Data does not contain any PII information.

Provider name: This a textual description of who is providing the data eg. "EDEKA".

7.1.2 Meta-data automatically assigned

Two additional fields will be automatically assigned to the dataset and data-source based on the delegated owner of the data-set.

- Data source id: A unique id for the data-source within the DIM.
- Provider (user) id: This is the identity of the data owner (or an authorised representative) who can make decisions on and agreements on data sharing.

7.1.3 Meta-data technical aspects

Some of the meta-data is more technical in nature, and refers to the collection, distribution and management of data itself.

URL to dataset. Where the data is provided as a linked data-set, this is the URL to fetch the corresponding data.

Access id, Access token. Necessary authentication information to access the data.

Output formats. Again selected from a defined series "json", "csv", "image-set", "sql"

Detailed description of the data. Where the data is structured, this is a more detailed description of the structure of the data. This allows use of columns of the dataset (perhaps merged with other data-sets), a description of each column present in the structured data is requested.

Where the data is not structured in any way, this includes the free text description of the data. It is intended that this data will be used as the basis for a textual description of the data available.

7.3 Data sources

In this third subsection, the main data sources to be considered are presented and analyzed.

7.3.1 Surveys (Quart de Poblet)

Information	Answer																							
Title of the data-source	“ALIMENTACIÓN SOSTENIBLE - CITIES2030”																							
Business sector	Municipality																							
Geographic-area covered by the data	Quart de Poblet, Valencia, Spain																							
Keywords	Sustainable food, consumption habits, survey, questionnaire																							
Production period	From: 2021-12-01 To: 2022-01-31																							
Sensitivity	☐ Anonymised - Anonymised data derived from personal data. ☐ Personal - Data set contains personal data. ☒ No personal data - Data does not contain any PII.																							
Provider name	Municipality of Quart de Poblet, Valencia, Spain																							
Formats	☐ json / REST ☐ image-set ☐ csv ☐ sql ☒ excel ☐ nosql ☐ Other																							
URL	Not applicable																							
Access id	Not applicable																							
Access token	Not applicable																							
Data source detailed description																								
<table><tr><th colspan="2">Column name</th><th rowspan="2">Description</th></tr><tr><th>Original (spanish)</th><th>English</th></tr><tr><td>ID</td><td>ID</td><td>An integer identifier within the document.</td></tr><tr><td>Hora de inicio</td><td>Starting time</td><td>Date and time of the beginning of the survey, in the format mm/dd/yy hh:MM:ss</td></tr><tr><td>Hora de finalización</td><td>End time</td><td>Date and time of the end of the survey, in the format mm/dd/yy hh:MM:ss</td></tr><tr><td>Sexo</td><td>Sex</td><td>Sex of the respondent (male/female)</td></tr><tr><td>Fecha de nacimiento</td><td>Birth date</td><td>Respondent’s birth date, in the format mm/dd/yy</td></tr><tr><td>Nivel de ingresos mensuales de la unidad familiar</td><td>Family’s income level</td><td>Total family’s income level, a value from the following list: < 900 </td></tr></table>		Column name		Description	Original (spanish)	English	ID	ID	An integer identifier within the document.	Hora de inicio	Starting time	Date and time of the beginning of the survey, in the format mm/dd/yy hh:MM:ss	Hora de finalización	End time	Date and time of the end of the survey, in the format mm/dd/yy hh:MM:ss	Sexo	Sex	Sex of the respondent (male/female)	Fecha de nacimiento	Birth date	Respondent’s birth date, in the format mm/dd/yy	Nivel de ingresos mensuales de la unidad familiar	Family’s income level	Total family’s income level, a value from the following list: < 900
Column name		Description																						
Original (spanish)	English																							
ID	ID	An integer identifier within the document.																						
Hora de inicio	Starting time	Date and time of the beginning of the survey, in the format mm/dd/yy hh:MM:ss																						
Hora de finalización	End time	Date and time of the end of the survey, in the format mm/dd/yy hh:MM:ss																						
Sexo	Sex	Sex of the respondent (male/female)																						
Fecha de nacimiento	Birth date	Respondent’s birth date, in the format mm/dd/yy																						
Nivel de ingresos mensuales de la unidad familiar	Family’s income level	Total family’s income level, a value from the following list: < 900																						

		● 900 - 1500 ● 1500 - 2000 ● 2000 - 2500 ● 2500 - 3000 ● 3000 - 3500 ● 3500 - 4000 ● 4000 - 4500 ● 4500 - 5000 ● > 5000
¿Dónde realiza la compra de fruta y verdura?	Where do you buy fruits and vegetables?	A value from the following list: ● Comercio local - local traders ● Supermercados - supermarkets ● No compro - Do not consume
¿Dónde realiza la compra de carne?	Where do you buy meat?	A value from the following list: ● Comercio local - local traders ● Supermercados - supermarkets ● No compro - Do not consume
¿Dónde realiza la compra de pescado?	Where do you buy fish?	A value from the following list: ● Comercio local - local traders ● Supermercados - supermarkets ● No compro - Do not consume
¿Dónde realiza la compra de pan y bollería?	Where do you buy bread and pastries?	A value from the following list: ● Comercio local - local traders ● Supermercados - supermarkets ● No compro - Do not consume
¿Consideras que tu dieta es saludable y sostenible?	Do you believe that your diet is healthy and sustainable?	A value from the following list: ● Mucho (A lot) ● Bastante (Good enough) ● Regular (Regular/Normal) ● Poco (Not much) ● Nada (Not at all)
¿En que piensas cuando te hablan de "alimentos sostenibles"?	What do you think about when talking about "sustainable food"?	One or more values from the following list: ● Bajo impacto medioambiental (Little environmental impact) ● Cadenas de suministros locales (Local supply chains) ● Evitar el uso de pesticidas y transgénicos (Avoiding usage of pesticides and GMOs)
¿Por qué no adoptas una alimentación más sostenible?	Why don't you adopt more sustainable food habits?	A value from the following list: ● No encuentro productos adecuados en mis lugares habituales de compra (I can't find suitable products in my usual places of purchase). ● Es demasiado caro (It is too expensive) ● Me falta un etiquetado claro (I lack clear labelling) ● Me falta información sobre como hacerlo (I lack information on how to do it) ● Considero que mi dieta es sostenible (I consider

		my diet to be sustainable)
Cuando compra comida, ¿Cuáles de los siguientes aspectos son los más importantes para usted?	When buying food, which of the following aspects are the most important for you?	Max. 3 items from the following list: ● Contenido nutricional (por ejemplo, la cantidad de vitaminas, proteínas, azúcares o grasas) - Nutritional content (e.g. amount of vitamins, proteins, sugars or fats) ● Origen del alimento - Origin of the food ● Coste - Cost ● Sabor - Taste ● Sus principios éticos y creencias (por ejemplo, el bienestar animal o sus preocupaciones sobre el medio ambiente) - Your ethical principles and beliefs (e.g., animal welfare or environmental concerns) ● Seguridad alimentaria (por ejemplo, si el consumo del alimento comporta algún riesgo) - Food safety (e.g., whether there are any risks involved in consuming the food)
De los siguientes problemas, cuáles le preocupan más cuando se trata de los alimentos?	Among the following problems, which disturb you the most with respect to food?	Max. 5 items from the following list: ● Los aditivos como colorantes, conservantes o aromas usados en alimentos o bebidas - Additives such as colorings, preservatives or flavourings used in food or beverages. ● Los contaminantes medioambientales en el pescado o la carne - Environmental contaminants in fish or meat ● Los residuos de pesticidas en los alimentos - Pesticide residues in food ● Los residuos antibióticos, hormonas o esteroides en la carne - Antibiotic, hormone, or steroid residues in meat. ● La higiene alimentaria - Food hygiene ● Los ingredientes modificados genéticamente en los alimentos o bebidas - Genetically modified ingredients in food or beverages ● Las enfermedades de los animales - Animal diseases ● Los microplásticos que se encuentran en los alimentos - Microplastics found in food ● La intoxicación alimentaria por bacterias - Food poisoning due to bacteria ● Las reacciones alérgicas a los alimentos o bebidas - Allergic reactions to food or beverages
¿Prestas atención al impacto de tus elecciones en el medio ambiente?	Do you pay attention to the impact of your consumption habits on the environment?	A value from the following list: ● Presto mucha atención - I pay close attention ● Algo de atención - Some attention ● Poca atención - Little attention ● No me importa nada - I don't care at all

7.3.2 Geospatial information

Here it is described the data source considered for the Geospatial information processing component, which is further described in D6.3.

Information	Answer
Title of the data-source	"OSM Points of interest"
Business sector	Food processing and market / sales
Geographic-area covered by the data	City level (10 cities included)
Keywords	Open Street Maps, Shops, Amenities, CRFS, nodes
Production period	From: 2022-06-01 To: 2022-08-23
Sensitivity	☐ Anonymised - Anonymised data derived from personal data. ☐ Personal - Data set contains personal data. ☒ No personal data - Data does not contain any PII.
Provider name	OpenStreetMaps
Formats	☐ json / REST ☐ image-set ☐ csv ☐ sql ☐ excel ☐ nosql ☐ Other
URL	Available through S2CP platform: https://cities2030.eu/single-click-crfs-platform/
Access id	No need
Access token	No need

7.3.2.1 Data source detailed description

For each city a couple of datasets have been downloaded:

1. Point-type set (OSM nodes), which represents the point-based elements of the establishments according to the categories present in Figure 1
2. Linestring type set (OSM Ways), which are presented in open or closed polygons, generally demarcating an area of an establishment, such as a supermarket.

All entities have some **mandatory information**, which is the <id> and the geographic coordinates of the place, as follows:

- Id: Identifier in the database
- Geometry-type: Point | Linestring
- Coordinates: Coordinates expressed in WSG84 standard.

Other properties are optional and are represented in the following table.

Deliverable D6.2

Table 4. Optional properties. Geospatial information

Property	Description	Property	Description
<i>addr</i>	full address, city, country	<i>min_age</i>	<Number>
<i>amenity</i>	restaurant bar pub, cafe supermarket etc	<i>name</i>	Name of shop or amenity
<i>brand</i>	Amenity brand name	<i>opening_hours:</i>	<Date Range>
<i>capacity</i>	<Number>	<i>outdoor_seating</i>	Yes no
<i>check_date</i>	<Date>	<i>payment</i>	Payment information: cash, contactless, credit, debit, electronic
<i>contact</i>	Contact information: email, facebook, fax, phone, tripadvisor, website, etc.	<i>phone</i>	<Phone number>
<i>created_by</i>	<Id>	<i>reservation:</i>	yes no
<i>cuisine:</i>	regional barbecue burger chicken fish pasta pizza tapas etc	<i>smoking:</i>	outside no inside
<i>delivery:</i>	Yes no	<i>takeaway:</i>	yes no
<i>description</i>	Description of shop or amenity	<i>toilets</i>	yes no, type of access: wheelchair Fee: yes no
<i>diet</i>	Gluten_free meat vegan vegetarian etc.	<i>website:</i>	<url>
<i>indoor_seating</i>	yes no	<i>website:menu</i>	Yes no
<i>internet_access:</i>	yes no with fee	<i>wheelchair:</i>	limited supported no

7.3.3 Social network information (Twitter)

Information	Answer
Title of the data-source	Twitter data
Business sector	Telecommunications
Geographic-area covered by the data	All world (can be controlled by the type of query)
Keywords	Social networks, twitter
Production period	Flexible, depends on the type of Twitter API being used.

Deliverable D6.2

	Limits: From: 2006-03-21 To: today
Sensitivity	☐ Anonymised - Anonymised data derived from personal data. ☐ Personal - Data set contains personal data. ☐ No personal data - Data does not contain any PII.
Provider name	E.g "EDUKA"
Formats	☐ json / REST ☐ image-set ☐ csv ☐ sql ☐ excel ☐ nosql ☐ Other
URL	Can be accessed via Twitter API https://developer.twitter.com/en/docs/twitter-api
Access id	Not applicable
Access token	Not applicable

7.3.3.1 Data source detailed description

Table 5. Twitter API returns the data in the following format (per tweet)

created_at	Date of the creation, in format yyyy-mm-dd hh:MM:ss
id	Unique identifier of the tweet, number format
id_str	Unique identifier of the tweet, string format
text	Displayed text of the tweet
display_text_range	Positions of the first and the last character of the tweet that are displayed
source	URL of the source
truncated	Boolean, if tweet is truncated
in_reply_to_status_id	ID of the status, to which this tweet is a reply, number format
in_reply_to_status_id_str	ID of the status, to which this tweet is a reply, string format
in_reply_to_user_id	ID of the user, to which this tweet is a reply, number format
in_reply_to_user_id_str	ID of the user, to which this tweet is a reply, string format
user	User description in JSON format, see below
geo	Geographical location of where the tweet was posted, if enabled by user
place	Description of the geographical location, if enabled by the user, see below
is_quote_status	Boolean, if a quote
extended_tweet	Fool information about the tween in JSON format, see below
quote_count	Number of times the tweet was quoted
reply_count	Number of times replies were posted to this tweet
retweet_count	Number of times the tweet was retweeted
favorite_count	Number of times the tweet was marked as favourite
retweeted	If this is a retweet
lang	Language of the tweet

Table 6. Description of the field 'user'

id	Unique identifier of the user, number format
id_str	Unique identifier of the user, string format
name	Name of the user
screen_name	Screen name of the user
location	User-defined location, a string which can contain any text
description	Self-description of the user
verified	If user is verified
followers_count	Number of followers
friends_count	Number of friends
favourites_count	Number of times marked favourite
statuses_count	Number of statuses
created_at	Date of the creation, in format yyyy-mm-dd hh:MM:ss
utc_offset	User's utc offset
time_zone	User's time zone
lang	language
profile_background_image_url	URL of the user's profile background image
profile_background_image_url_https	URL of the user's profile background image (HTTPS format)
profile_image_url	URL of the user's profile image
profile_image_url_https	URL of the user's profile image (HTTPS format)

Table 7. Description of the field 'place'

id	ID
place_type	Type of the place, for example city, village etc.
name	Name of the place, for example 'Coyoacán'
full_name	Full name of the place, for example 'Coyoacán, Distrito Federal'
country_code	Country code
country	Country full name
bounding_box	Geographical coordinates of a geometric figure that covers the place on a map

Table 8. Description of the field 'extended_tweet'

full_text	Full text of the tweet
entities	Information about hashtags, urls, and users mentioned in the tweet
media	Information about the media used in the tweet, such as photos and videos. This field contains the type of the media, its description such as size, and access URLs.

7.3.4 Sensor real-time data (Arganda)

Below, it is shown the information that is defined for the capturing of real time information in Arganda Lab. Although the system is still in development phase, Lab responsible partners have facilitated the data model

Deliverable D6.2

Project 'cities2030' | H2020 ID | 101000640 | 'Co-creating resilient and sustainable food systems towards FOOD2030' | www.cities2030.eu
that they have defined, for the capturing of information generated by the censoring infrastructure that is being installed.

Information	Answer
Title of the data-source	Realtime data example of food processing company operation
Business sector	Food processing
Geographic-area covered by the data	Arganda
Keywords	Sensor information, food processing, temperature, humidity, air quality
Production period	To be determined
Sensitivity	☐ Anonymised - Anonymised data derived from personal data. ☐ Personal - Data set contains personal data. ☒ No personal data - Data does not contain any PII.
Provider name	Universidad Politécnica de Madrid
Formats	☐ json / REST ☐ image-set ☐ csv ☐ sql ☐ excel ☐ nosql ☒ Other: jsonld
URL	No available yet. It will be available through S2CP platform web page: https://cities2030.eu/single-click-crfs-platform/
Access id	Not needed
Access token	Not needed

7.3.4.1 Data source detailed description

Data will be captured by smart sensor nodes (powered by an Arduino Mega architecture with an Ethernet W5100 shield), combining various sensors:

- DTH-11 module to measure environmental temperature and humidity
- CCS811 module to monitor the air quality (equivalent CO₂ -eCO₂- and total volatile organic compounds - TVOC-).

Data structure will be expressed in json-ld format, by following DEMETER's Agricultural Information Model¹⁸⁴. The structure is described as follows:

Table 9. Properties for real-time data

Resource	Property	Value
Id	@type	urn:upm:demeter:p52:node
Position	http://www.opengis.net/ont/sf#Point	geosparql#wktLiteral
urn:upm:demeter:p52:node	@type	Observation
urn:upm:demeter:p52:node	resultTime	<Date>
urn:upm:demeter:p52:node	temperature	http://qudt.org/vocab/unit/DEG_C

¹⁸⁴ <https://github.com/rapw3k/DEMETER>

urn:upm:demeter:p52:node	humidity	http://qudt.org/vocab/unit/PERCENT_RH
urn:upm:demeter:p52:node	co2	http://qudt.org/vocab/unit/PPM
urn:upm:demeter:p52:node	tvoc	http://qudt.org/vocab/unit/Milligram-per-M3

7.4 General analysis of relevant external and internal data sources

From the above use-case analysis, three different types of data-source can be derived.

- **Structured:** This is a data-source that inherently provides some structure to the data supplied. Examples of this type of data-source include CSV files, Excel worksheets and Relational databases.
- **Unstructured:** This data-source is focused on documents. The significant difference is that the documents are assumed to be opaque with no predefined structure. An example could be a text document containing customer feedback for a supermarket chain.
- **Linked:** This is where the data-source itself is already object based. For example, a set of pictures used to promote a given city. Here the data is a set of images available for use. In these cases it is sufficient to link to the data-items for most use cases. In some cases, these data objects have accompanying meta-data that is in a structured format. Eg. EXIF tags on photos.

8 Conclusions and next steps

This document aimed to provide guidance in two areas. First, this document forms a comprehensive review of the state-of-the-art for data mining and other data processing activities. Where appropriate a comparison between techniques and tools has been provided to help guide the implementation work. Some of the techniques described are currently not very useful for the use cases Cities2030 proposes and therefore have been de-prioritised for implementation activities.

A second aspect is the formal definition of the components that need to be developed or augmented for use on Cities2030. These components were identified by consultation with the Labs on what specific tools would be needed to support their work. In all cases, the interaction points between these components have been identified and documented. After careful analysis some design decisions have been made to support the collection, discovery and combination of data-sources to support the use cases identified by the Labs.

Using careful design, hooks have been provided to allow Data security and governance issues to be addressed by additional components provided by T6.4. This is supported by the identification of data-source meta-data which can be used to add value to the data and clearly identify data-ownership and promote a data marketplace that benefits the producers (and owners) of the data.

During the next phase of the project, the data mining and open data repository will be implemented, augmented and released for use by the associated Labs. The platform will provide tools to aid the management and collection of data sources, In addition it will provide tools to aid the discovery of data-sets and data-sources; promoting reuse of data-sets; and enabling new combinations of data to provide a data-driven analysis platform that can provide the insight requested by the Labs.

These components will be refined further based on the feedback received from the Labs using these components. An update on these improvements will be provided in a revised Deliverable as the next phase executes.

Annex I - Quart de Poblet use case

Resources:

1. Survey conducted by Quart de Poblet, Valencia, Spain.
2. Analysis of social network data (Twitter)

The aim of this case study is to understand the opinion of consumers about sustainable food practices and to what extent these practices are adapted in their daily life. To conduct this study, we used data from a survey done by Quart de Poblet in Valencia, Spain and we performed analysis of data from Twitter. The survey contains answers from 189 respondents. The answers were collected during the period 28.12.2021 – 14.01.2022.

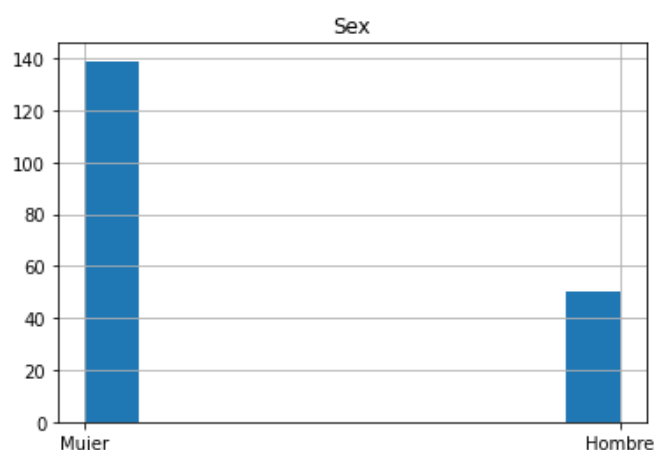
Survey analysis

General information about participants

Sex: Most participants are women (74%)

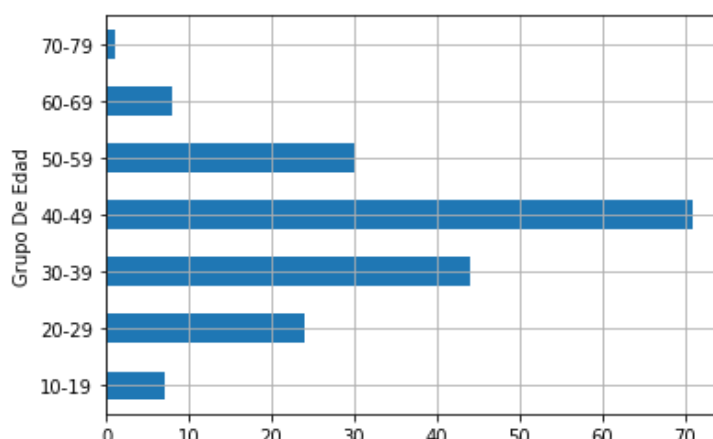
26 % of participants are men (50)

74 % of participants are women (139)



Age distribution:

Grupo De Edad	count	%
10-19	73	38.6
20-29	241	126.5
30-39	442	233.3
40-49	713	375.2
50-59	301	159.3
60-69	84	44.4
70-79	10	5.3

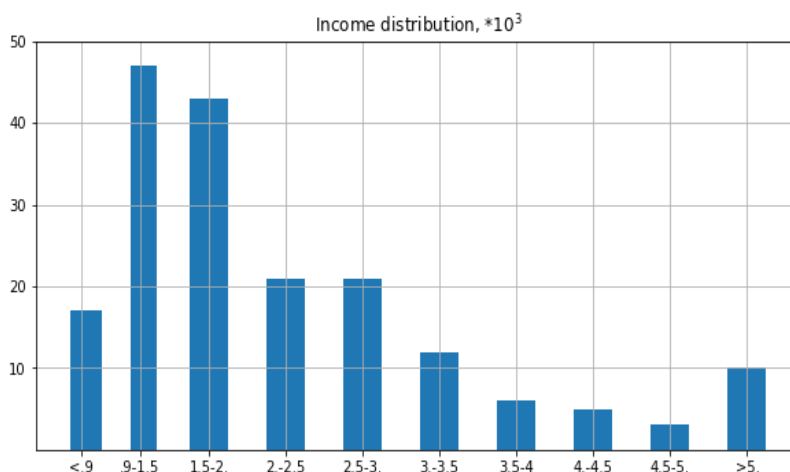


Almost 40% of respondents are between 40 and 50 years old. In general, the age of participants looks normally distributed around the mean value 40-50 with a shift towards the younger people. This indicates high inclusion of the younger generation into the survey.

Income distribution:

Nivel de ingresos mensuales	num%
< 900	179.2
900 - 1500	4725.4
1500 - 2000	4323.2
2000 - 2500	2111.4
2500 - 3000	2111.4
3000 - 3500	126.5
3500 - 4000	63.2
4000 - 4500	52.7
4500 - 5000	31.6
> 5000	105.4

Almost 63% of respondents have an income level between 900 and 3000 euros per month. 9% of respondents have the lowest income level below 900 euros per month. This group of people is likely to adopt less sustainable consuming habits due to their low income. 19.4% of respondents have relatively higher income values above 3000 euros per month.



Summary

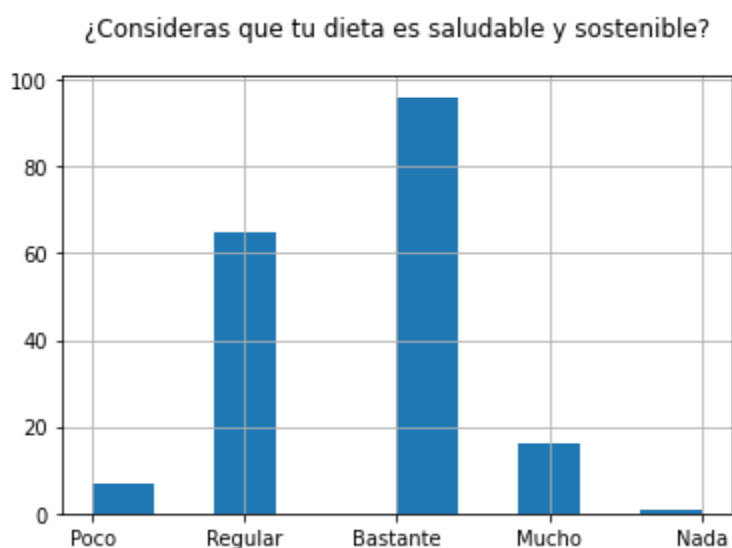
The survey covers various strata of population with the biases towards females and younger age groups. This seems reasonable for the current study, as women are more often responsible for the household management and younger generations are more active.

Participants' awareness

Do you consider your diet to be healthy and sustainable?

Around 96% of respondents consider their diet to be healthy and sustainable more or less.

- 1% (1) of participants do not have a healthy and sustainable diet at all
- 4% (7) of participants have a less healthy and sustainable diet
- 35% (65) of participants have a moderately healthy and sustainable diet
- 52% (96) of participants have a fairly healthy and sustainable diet
- 9% (16) of participants have a very healthy and sustainable diet



What do you think about when discussing "sustainable food"?

(several answers per participant are possible)

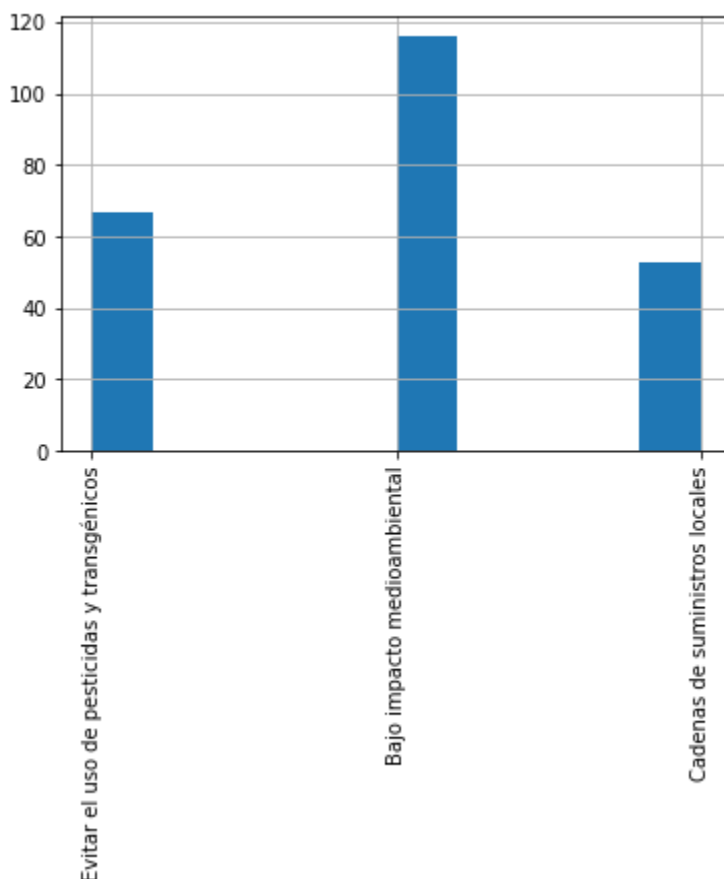
A high proportion of participants (63%) consider sustainable food as a way to lower the environmental impact.

36% (67) of participants think about sustainable food as a way to avoid the use of pesticides and GMOs

63% (116) of participants think about sustainable food as a way to lower environmental impact

28% (53) of participants believe that sustainable food has a direct relation with local supply chains

¿En que piensas cuando te hablan de "alimentos sostenibles"?



Why don't you adopt more sustainable consumption habits?

Only 13% of respondents consider their diet to be sustainable enough. As for the rest 87%, some of the other reasons stop them from adopting more sustainable consumption habits: 26% need more information on how to do it, 24% believe that a sustainable diet is too expensive, 19% require clear labelling and 18% cannot find suitable products in their usual shopping places.

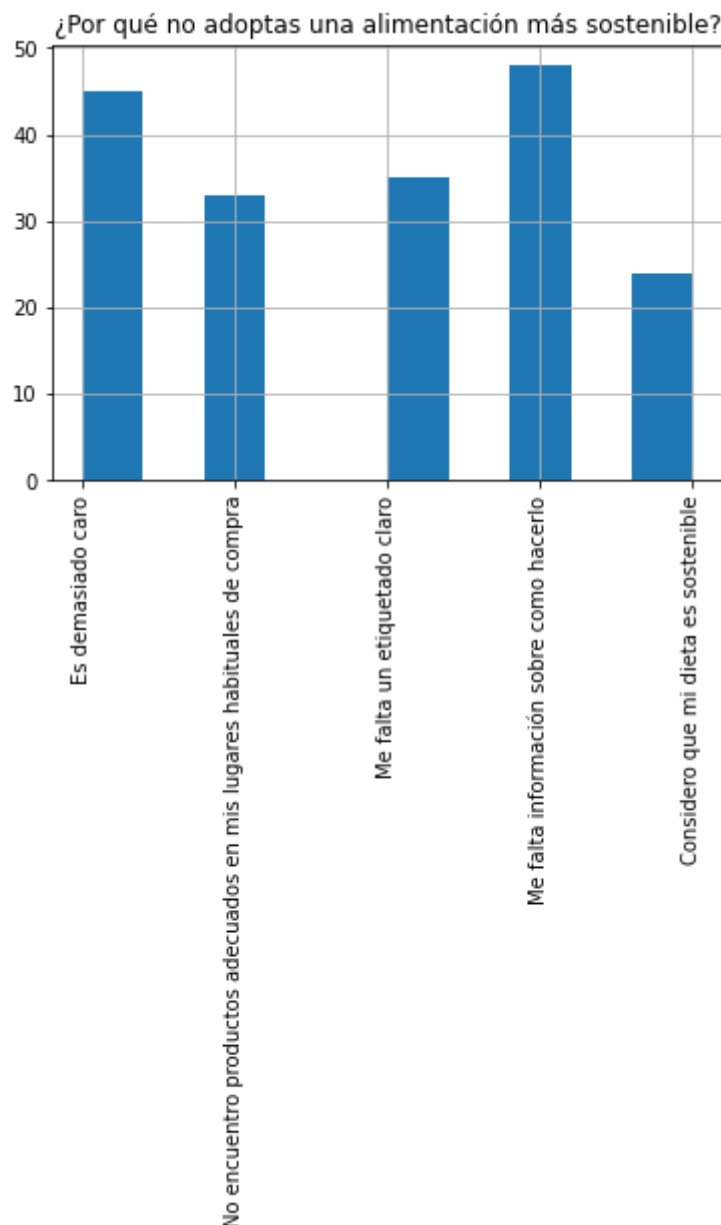
24% (45) of participants do not follow a sustainable diet because it is too expensive

18% (33) of participants do not follow a sustainable diet because they cannot find suitable products in their usual shopping places

19% (35) of participants do not follow a sustainable diet because they need a clear labelling

26% (48) of participants do not follow a sustainable diet because they need information on how to do it

13% (24) of participants have a sustainable diet



What are the most important aspects for you when you are buying food?

(multiple answers are possible)

The following aspects are important:

- nutrition content (for 60% of respondents)
- price (50%)
- taste (43%)
- safety (43%)
- origin (42%)

As we can see, safety and origin, although being important, are less often important than nutrition content and price.

Only 12% of respondents believe that they follow their ethical principles with respect to consumption habits.

for 60% (111) of participants, the nutrition content is important

for 42% (78) of participants, the origin of food is important

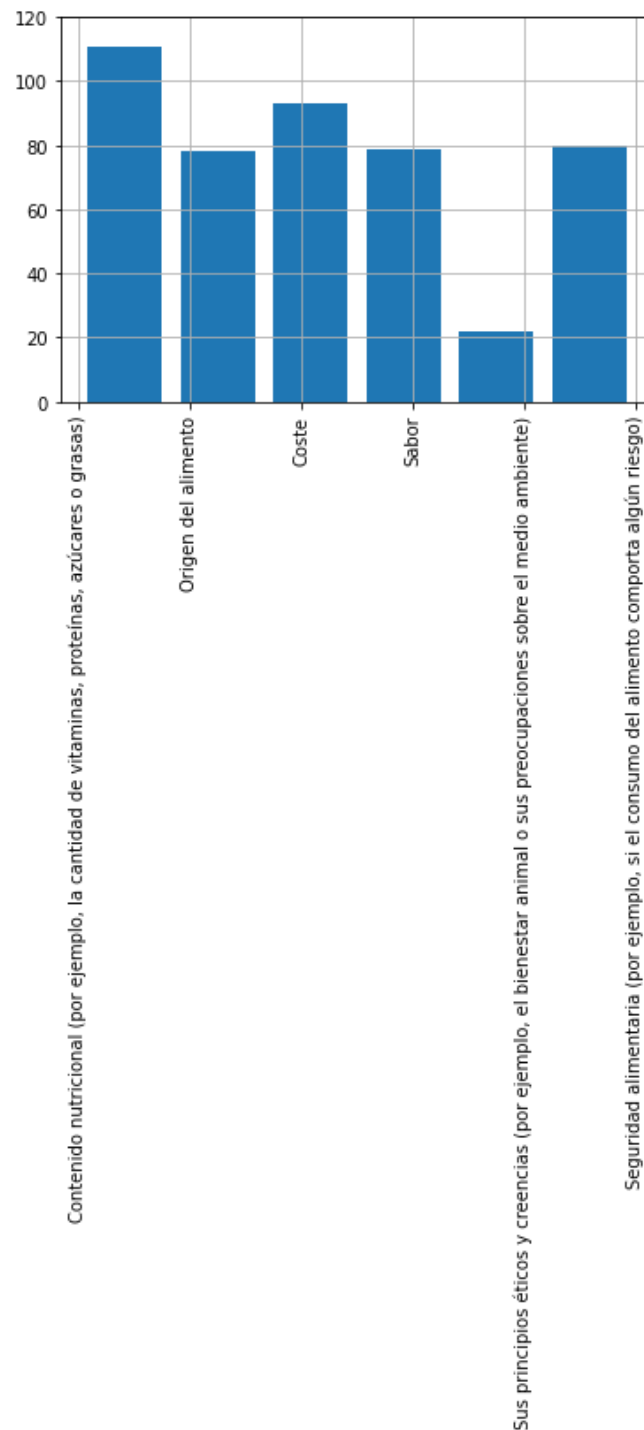
for 50% (93) of participants, the price of food is important

for 43% (79) of participants, the taste of food is important

12% (22) of participants follow their ethical principles and beliefs when it comes to buying food

43% (80) of participants care about food safety

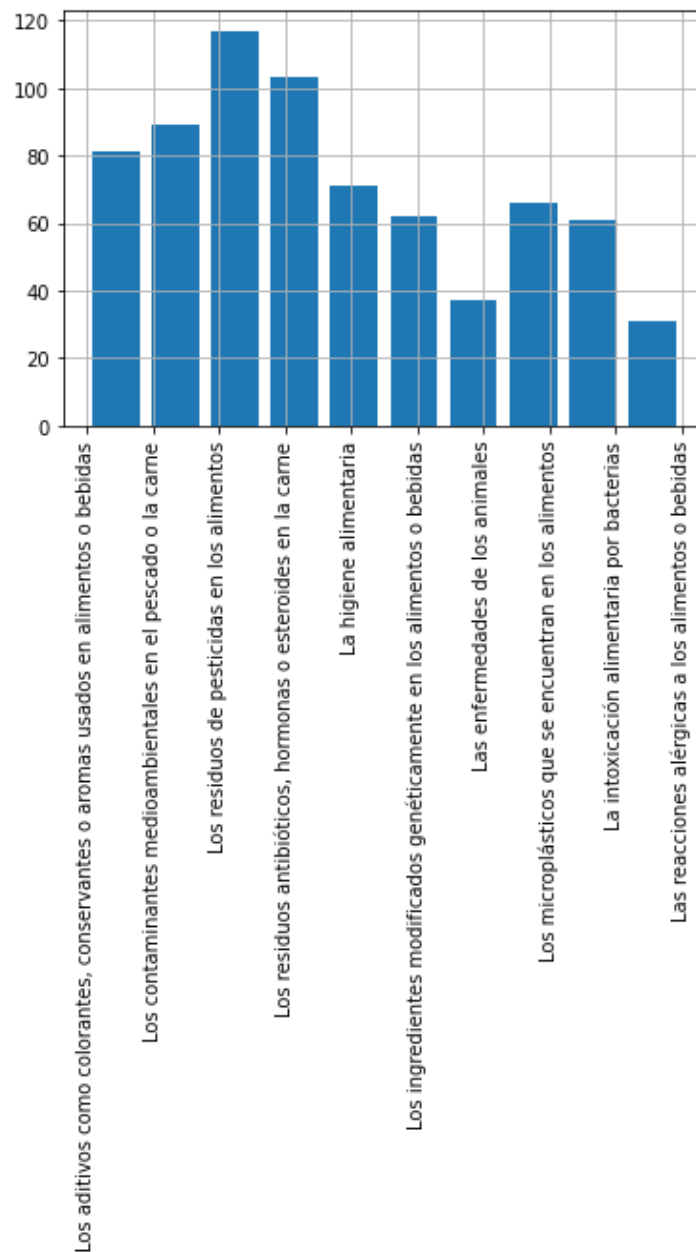
Quando compra comida, ¿Cuáles de los siguientes aspectos son los más importantes para usted? (Max. 3 respuestas)



Among the following problems, what are those that bother you the most when concerning food?
 (multiple answers are possible)

The following problems are important for the consumers:

- pesticide residues in food (63%)
- antibiotic, hormone, or steroid residues in meat (56%)
- environmental contaminants in fish or meat (48%)
- additives such as colorants, preservatives, or flavorings (44%)
- food hygiene (38%)
- microplastics in food (36%)
- bacterial food poisoning (33%)
- animal diseases (20%)
- allergic reactions to food or drinks (17%)

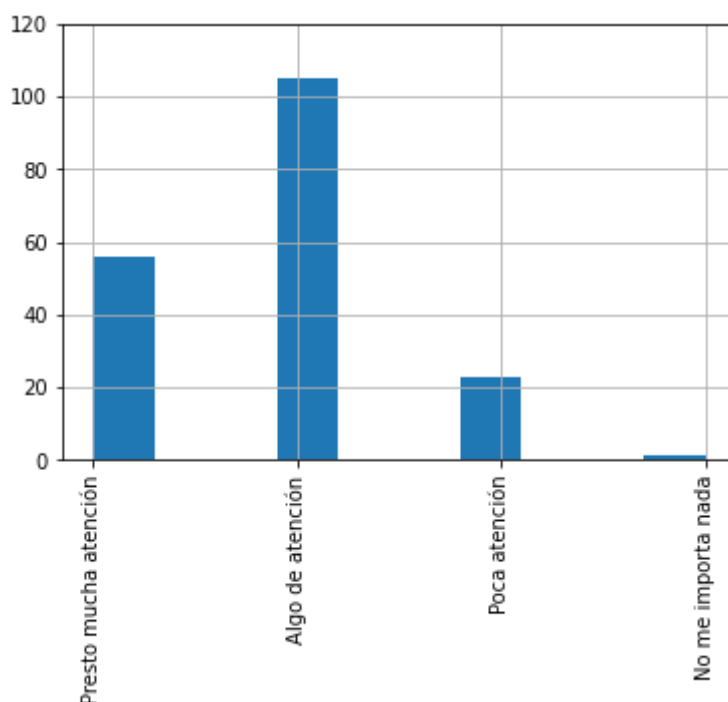


44% (81) of participants are concerned about additives such as colorants, preservatives, or flavorings
 48% (89) of participants are concerned about environmental contaminants in fish or meat
 63% (117) of participants are concerned about pesticide residues in food
 56% (103) of participants are concerned about antibiotic, hormone, or steroid residues in meat
 38% (71) of participants are concerned about food hygiene
 34% (62) of participants are concerned about genetically modified ingredients in food or drinks
 20% (37) of participants are concerned about animal diseases
 36% (66) of participants are concerned about microplastics in food
 33% (61) of participants are concerned about bacterial food poisoning
 17% (31) of participants are concerned about allergic reactions to food or drinks

Do you pay attention to the impact of your choices on the environment?

In general, most of the participants (86%) do care about the impact on the environment with 30% paying a lot of attention and 56% paying some attention. 14% pay no or little attention this this issue.

30% (56) of participants pay a lot of attention to the impact of your choices on the environment
 56% (105) of participants pay some attention to the impact of your choices on the environment
 12% (23) of participants pay a little attention to the impact of your choices on the environment
 1% (1) of participants do not care about the impact of their choices on the environment



Summary

1. Impact on environment:

- Most of the participants (86%) do care about the impact on the environment with 30% paying a lot of attention and 56% paying some attention. 14% pay no or little attention this this issue.

- A high proportion of participants (63%) consider sustainable food as a way to lower the environmental impact.

2. Sustainability:

Around 96% of respondents consider their diet to be healthy and sustainable more or less. However, only 12% of respondent believe that they follow their ethical principles with respect to consumption habits. Also, only 13% of respondents consider their habits to be sustainable enough. Others report reasons that stop them from adopting more sustainable consumption habits:

1. lack of information on how to do it
2. lack of clear labelling
3. absence of sustainable products in their usual shopping places
4. sustainable diet is too expensive

The first 2 reported reasons clearly demonstrate the possibility to improve the situation by providing more information.

3. What are the respondents concerned about?

1. aspects:

- nutrition content (for 60% of respondents)
- price (50%)
- taste (43%)
- safety (43%)
- origin (42%)

2. Problems:

- pesticide residues in food (63%)
- antibiotic, hormone, or steroid residues in meat (56%)
- environmental contaminants in fish or meat (48%)
- additives such as colorants, preservatives, or flavorings (44%)
- food hygiene (38%)
- microplastics in food (36%)
- bacterial food poisoning (33%)
- animal diseases (20%)
- allergic reactions to food or drinks (17%)

Analysis of consuming habits

Source of Bread and pastries

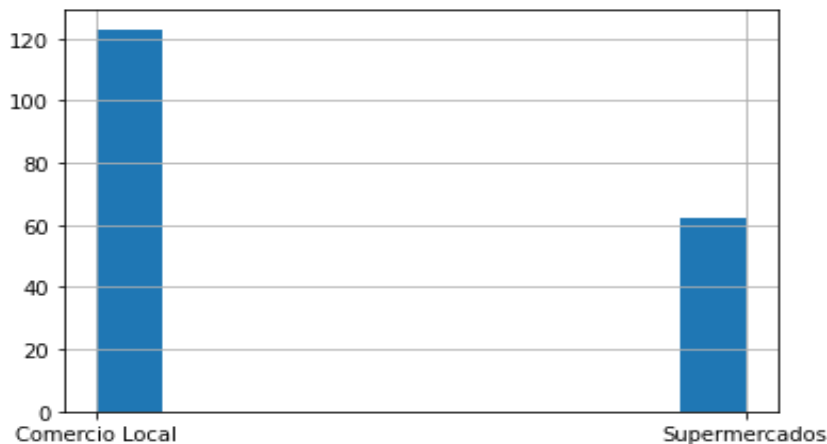
Almost 2/3 of respondents prefer buying bread and pastries from local markets.

66% (123) of participants buy pastries from Local Trades

34% (62) of participants buy pastries from Supermarkets



¿Dónde realiza la compra de pan y bollería?



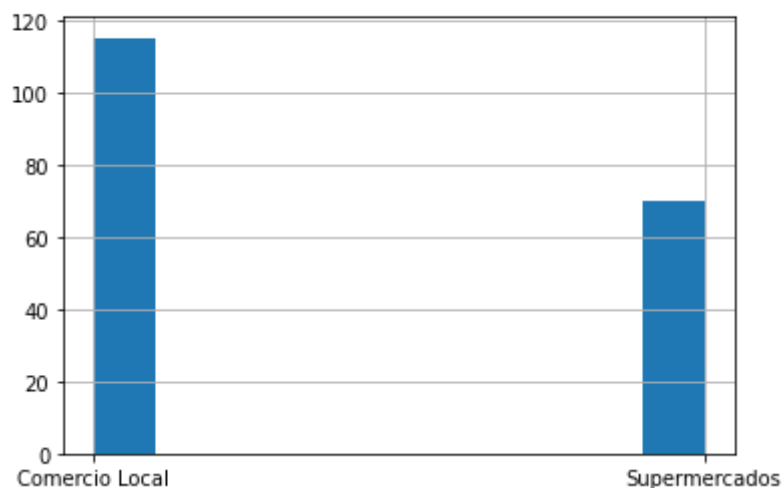
Source of fruits and vegetables

Preference of buying vegetables and fruits from local markets is also high. However, the fraction of respondents who choose it falls from 66% to 62%.

62% (115) of participants buy fruits and vegetables from Local Trades

38% (70) of participants buy fruits and vegetables from Supermarket

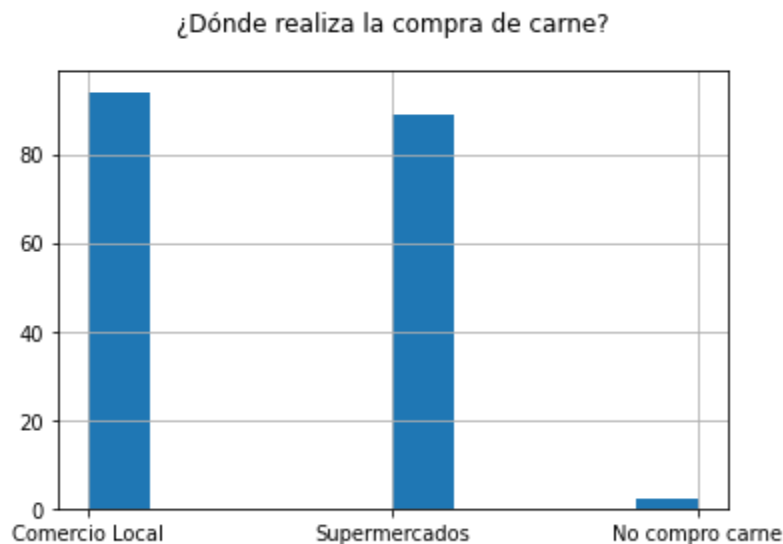
¿Dónde realiza la compra de fruta y verdura?



Source of meat

Moving to meat, the fraction of respondents preferring local markets falls even further. Now, it is only 51%. 1% of respondents reported not to buy meat at all.

51% (94) of participants buy meat from Local Trades
 48% (89) of participants buy meat from Supermarkets
 1% (2) of participants do not buy meat at all

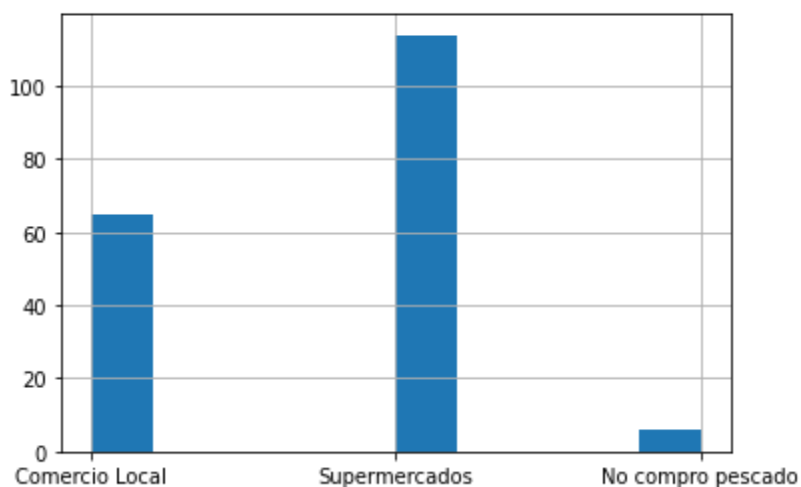


Source of Fish

Contrarily to other products, the fish is most often bought from supermarkets (62% or almost 2/3 of respondents). 3% of respondents do not buy fish at all.

35% (65) of participants buy fish from Local Trades
 62% (114) of participants buy fish from Supermarkets
 3% (6) of participants do not buy fish at all

¿Dónde realiza la compra de pescado?



Summary

We can observe that in most of the cases buying from local traders is preferred. This trend is especially prominent in case of bread and pastries and fruits and vegetables. When buying meat, then numbers are distributed almost evenly (51% prefer local markets and 48% prefer supermarkets). The clear outlier from this pattern is the consumption of fish with an inverse tendency (2/3 of respondents prefer to buy fish from supermarkets).

From the section "Analysis of consuming habits", we also know that 48% of respondents are concerned about environmental contaminants in fish and meat. In particular, this applies to 47% of those respondents, who buy fish from Supermarkets. These respondents might be willing to change their habits if an appropriate option is suggested.

In the next section of the case-study, we try to further investigate the choices of consumers when buying fish.

Association rules mining

To further investigate why fish is more often bought from supermarkets than the other products, we performed association rule mining on the available dataset. We searched for rules with high confidence ($\geq 70\%$) and reliable lift (≥ 1.1). The support was not required to be very large ($\geq 10\%$). We were specifically interested in rules that have '¿Dónde realiza la compra de pescado? = Comercio Local' as the consequent to see what a stimulus for this choice can be. From the available dataset, we got 19 such rules like the one below:

'Sexo = Mujer'
 &
 '¿Dónde realiza la compra de carne? = Comercio Local'
 &
 '¿Dónde realiza la compra de pan y bollería? = Comercio Local'
 &
 '¿En que piensas cuando te hablan de "alimentos sostenibles"? = Evitar el uso de pesticidas y transgénicos'

[Symbol]

'¿Dónde realiza la compra de pescado? = Comercio Local'

supp = 0.103

conf = 0.704

lift = 2.0

This rule reads as follows:

If <sex is female> and <meat and bread/pastries are bought from local producers> and <sustainable food practices mean avoiding the use of pesticides and transgenics>, then <the fish is bought from local producer>.

This rule applies to 10.3% of the respondents. Confidence of 70.4% means that in 70.4% of the cases when the if-cause is true, the fish is bought from local producers. The fact that in the survey there are only 3 possible options for fish consumption ('local producer', 'supermarkets' and 'do not consume fish'), means that in the other 30% of the cases the fish is either not bought at all or is bought from a supermarket. Finally, the value of lift = 2.0 ($2.0 > 1$) indicates that this rule does not reflect a spurious association but a piece of reliable information.

It is interesting to notice, that among the 19 discovered rules, 16 rules (84%) contain in the if-clause the fact that meat is also bought from the local market ('¿Dónde realiza la compra de carne? = Comercio Local'). The other 3 rules contain in the if-clause the fact of buying vegetables and bread/pastries from local market ('¿Dónde realiza la compra de fruta y verdura?=Comercio Local', '¿Dónde realiza la compra de pan y bollería?=Comercio Local'). This means that buying other products from the local market is a necessary condition for buying fish from local market.

At the same time, if we perform a similar analysis for "buying meat at local markets" ('¿Dónde realiza la compra de carne? = Comercio Local'), we can find many rules that have no conditions on buying other products in the if-clause, like the one below:

'Sexo = Mujer'

&

'Grupo De Edad = 40-49'

&

'¿Consideras que tu dieta es saludable y sostenible? = Bastante'

[Symbol]

'¿Dónde realiza la compra de carne? = Comercio Local'

supp = 0.114

conf = 0.724

lift = 1.425

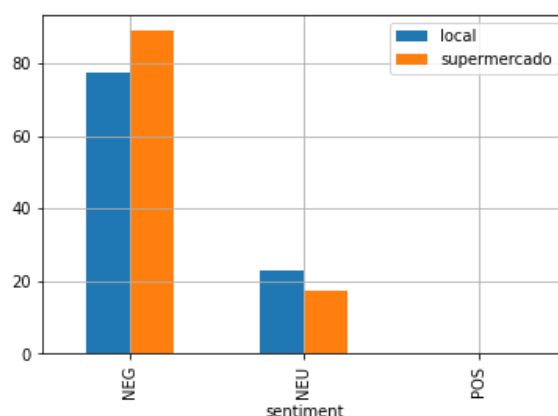
This rule reads as follows:

If <sex=female> and <age_group = 40-49> and <considers your diet to be healthy and sustainable = sufficiently>, then <meat is bought from local market>.

This observation once again shows some particularities with consumption habits for fish.

Analysing opinions through social media data mining (Twitter data analysis)

To better understand what drives people's choices to buy fish from local producers or from supermarkets, we perform analysis of data from social networks, namely twitter. We collected tweets corresponding to the queries "pescado local" (fish local) and "pescado supermercado" (fish supermarket) for a 1 month period 05.04 – 05.05.2022. Totally, we got 207 tweets for the query "pescado local" and 45 for the query "pescado supermercado". From these numbers we can already observe that the topic of buying local fish attracts more attentions and is discussed more often on this social media platform. On the next step, we performed sentiment analysis of the associated tweets.



Data for the plot:

"pescado local", tot=207			"pescado supermercado", tot=45		
Negative	160	77.3%	Negative	40	88.9%
Neutral	47	22.7%	Neutral	5	17.1%
Positive	0	0	Positive		

As we can see, for both search queries the sentiment of discussion is mostly negative corresponding to 77.3% for "pescado local" and 88.9% for "pescado supermercado". The second query more often has tweets with negative connotations. This indicates that this is a hot topic with active and emotional discussion. The absence of tweets with positive sentiment means that mostly the negative aspects of both options are being discussed.

Below we analyse some of the negative tweets corresponding to each query

text in Spanish	text in English	Aspect the people are discussing
"pescado local"		

Hoy en otro capítulo de por qué hay que apoyar el comercio local, cuando te pesan la carne/pescado y te cobran eso, pero luego al preparártelo y quitarte ⅓ del peso sigues pagando por la cantidad total.	Today in another chapter on why you have to support local businesses, when they weigh your meat/fish and charge you that, but then when you prepare it and take ⅓ of the weight off, you continue to pay for the full amount.	price
Yo fui este año y barato no es. Sale relativo, por qué los paquetes no son todo incluido, y yo soy gastona va, pero estando allá a una le dan ganas de probar la comida local (sopas marineras, pescado, ceviche) y eso no es barato	I went this year and it is not cheap. It's relative, because the packages aren't all inclusive, and I'm a big spender, but being there makes you want to try the local food (seafood soups, fish, ceviche) and that's not cheap	price
Me gusta ser esa persona que comparte el video del pescado que tiene basura en su estómago y en mi historia siguiente participar por una tabla de 50 sushis ...	I like to be that person who shares the video of the fish that has garbage in its stomach and in my next story participate for a table of 50 sushi ...	pollution
... Hay un local importador de pescado en caballito que vende solo pescado freezado y te pone en el cartel que la única manera de matar una bacteria mortal que vive en el salmon es teniéndola un par de días por debajo de 15 grados bajo cero.	... There is a local fish importer that sells only frozen fish and puts you on the sign that the only way to kill a deadly bacteria that lives in salmon is to keep it below 15 degrees below zero for a couple of days.	quality
"pescado supermercado"		
Ahora entiendo a mi abuela cuando de pequeño le preguntaba que por qué compraba la carne en la carnicería, la fruta en la frutería y el pescado en la pescadería, teniéndolo todo en el supermercado.	Now I understand my grandmother when, as a child, she asked her why she bought meat at the butcher, fruit at the greengrocer and fish at the fishmonger, having everything in the supermarket.	Usual places of buying products, convenience

No entiendo cómo logran tener caseros/tiendas de cada cosa: para las frutas uno, los huevos otro, el pan otro, la carne otro, el pescado, los artículos de limpieza, y así. Yo no tengo cabeza, ni siquiera por lograr un ahorro. Compro absolutamente todo de una en el supermercado.	I don't understand how they manage to have homemade/shops for everything: one for fruit, another for eggs, another for bread, another for meat, fish, cleaning supplies, and so on. I have no head, not even to save money. I buy absolutely everything in one at the supermarket.	Usual places of buying products, convenience
Cada día se come peor. Se cocina menos y eso siempre significa comer peor. Haciendo la compra en cualquier supermercado se ve rápido en las cestas de la compra. Pero precio de pescado, frutas y verduras al alza y pocas ganas/tiempo de cocinar	Every day eating worse. You cook less and that always means eating worse. When shopping in any supermarket, you can see quickly in the shopping baskets. But price of fish, fruits and vegetables on the rise and little desire/time to cook	Lifestyle
Más de 300 kilos de carne, pescado y pollo, fueron incautados por las autoridades sanitarias en un supermercado ... Los productos se encontraban en descomposición y mal refrigerados según ...	More than 300 kilos of meat, fish and chicken were seized by health authorities in a supermarket ... The products were decomposing and poorly refrigerated according to ...	Food storage

Analysis of the text of tweets and their sentiments indicates that both solutions “buying local” and “buying from supermarkets” are not satisfactory at the current stage for the consumers. Both have their negative sides, such as high price, quality, pollution, relation to the current lifestyle. More efforts must be made from the side of the authorities, “local consumption of fish” being more preferable than “buying fish from the supermarket”.

We can also see that the drawback of each approach discovered from twitter analysis corresponds to those highlighted by the respondents in the survey, see Summary for Section “Participants’ awareness”. This means that in general analysis of social media is a reliable source of useful information.

Annex II - Arganda use case

CODAN is a company focused on bakery products whose main production infrastructure is placed in Arganda (Madrid). This infrastructure includes two types of ovens: electrical and combustion. While electrical ovens are very safe, modern, and well isolated, combustion ovens are legacy systems with several inconveniences such as uncontrolled heat transmission to the environments, CO₂ production and other organic gases (potentially toxic).

In order to monitor these variables, CODAN has deployed some sensor nodes controlling some parameters: environmental temperature, environmental humidity, eCO₂ emitted to the atmosphere and volatile components emitted to the atmosphere. These values are obtained each 30 seconds, so the amount of available data is relevant.

CODAN would like “to share” this information with some other agents but controlling at every moment who can access the data, with which purpose and which permissions. These permissions need to be managed dynamically. These data are very sensitive for them, as they reveal critical information such as their production schedule.

CODAN defines the following permissions:

- Visualization: The agent can see data with full precision and timestamps regarding a giving time interval
- Anonymous visualization: The agent can see data but with a limited precision and without any time reference, regarding a giving time interval
- Manipulation: The agent is allowed to process and manipulate data (regarding a giving time interval), but it cannot store the raw information nor the results, which must be also shared with CODAN
- Property: The agent can store, visualize and manipulate data regarding a giving time interval. The results of any manipulation must be also shared with CODAN.

Four kinds of agents are considered by CODAN:

- Agent#1 (Cloud companies and digital service providers): CODAN needs to visualize data at real-time, so they can made decisions in a very efficient and fast manner. Besides, enriched digital services are also essential. These companies should be only provided with “manipulation” permission.
 - Agent#2 (Research institutions): CODAN collaborates with different research institutions to support the Industry 4.0 revolution. In this context, data-oriented solutions are required and CODAN’s information may be essential. CODAN is open to share data with research institutions with “property” permission, so researchers can work easily.
 - Agent#3 (Government departments): Periodically, CODAN must face administrative evaluations. In this case, government departments can access historical data with “visualization” permission to perform the required checks.
 - Agent#4 (Social agencies, consumers, etc.): Finally, CODAN is supporting a higher transparency in the economy and the food supply chain, proving final customer, social agencies, etc. with data about how their products are manufactured, the working conditions of people, the environmental impact, etc. Thus, all these agents might access CODAN’s data but with “anonymous visualization” permissions
1. Defined set of Roles for access control
 2. Defined set of Agents per Role
 3. Defined and accessible Data Store and REST API exposing JSON-LD

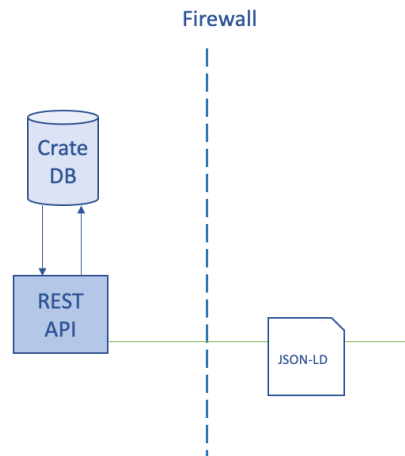


Figure 26. ARGANDA Legacy Data Infrastructure

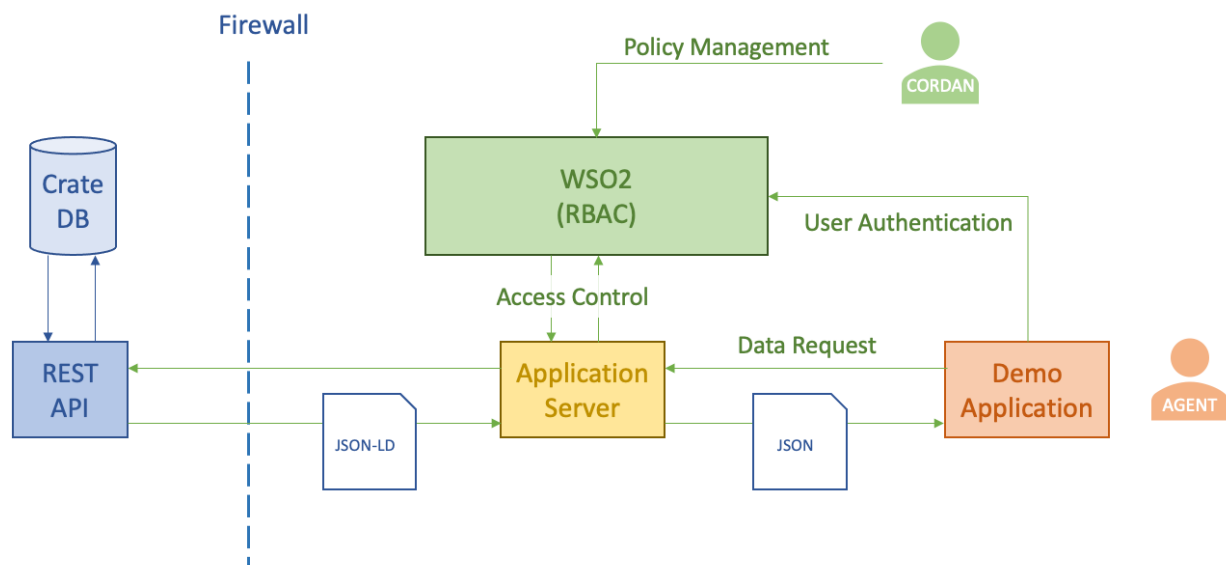


Figure 27. ARGANDA Use Case proposed solution