

Design and development of a unified framework for the anonymization, analysis, visualization and exploration of big data acquired from digital market places

Emmanouil Adamakis

Thesis submitted in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes, Heraklion, GR-70013, Greece

Thesis Advisor: Prof. *Constantine Stephanidis*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).

UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

**Design and development of a unified framework for the anonymization, analysis,
visualization and exploration of big data acquired from digital market places**

Thesis submitted by
Emmanouil Adamakis
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author:



Emmanouil Adamakis

Committee approvals:

**KONSTANTINOS
STEFANIDIS**

Digitally signed by
KONSTANTINOS STEFANIDIS
Date: 2022.01.07 17:27:41 +02'00'

Constantine Stephanidis
Professor, Thesis Advisor

**KONSTANTINOS
MAGOUTIS**

Digitally signed by
KONSTANTINOS MAGOUTIS
Date: 2022.01.07 23:43:35 +02'00'

Kostas Magoutis
Associate Professor, Committee Member



Margherita Antona
Principal Researcher, Committee Member

Departmental approval:

**Polyvios
Pratikakis**

Digitally signed by Polyvios
Pratikakis
Date: 2022.01.23 09:56:53
+02'00'

Polyvios Pratikakis
Associate Professor, Director of Graduate Studies

Heraklion, December 2021

Design and development of a unified framework for the anonymization, analysis, visualization and exploration of big data acquired from digital market places

Abstract

In today's data-driven world, data interchange plays a pivotal role in our daily lives. Every digital transaction, from the simplest to the most complex, requires data exchanging between the parties involved. From individuals and small businesses to large corporations, organizations, and governments all store, process and exchange data. This situation, over time, has led to the accumulation of large volumes of data, called Big Data. With the emergence of Big Data, it became apparent that there were numerous opportunities in terms of their analysis and the information results (insights) of such analyses, which could be highly beneficial to the data processors' goals. Of great assistance at improving the outcomes of such analyses was also identified to be the enrichment and correlation of existing internal datasets with datasets acquired from external sources. Obtaining third-party datasets used to entail approaching specific data owners directly; however, with the emergence of digital data market places in recent years, this situation has begun to change.

Until recently, data exchanges were carried out with little to no regard for privacy or the protection of personal data. Recent legislative developments, such as the European Union's GDPR data protection laws, have prompted many data providers and consumers to seek solutions for both protecting individuals' privacy and assessing the privacy risks of the datasets under their management. Following these developments, any data disclosure has to employ some form of data sanitization prior to release, in order to protect the privacy of individuals' sensitive information. Anonymization of data is an example of such a sanitization process, and it involves the deduction or transformation of data in a privacy-preserving manner in order to achieve a certain level of anonymity. One of the most difficult aspects of any anonymization process is striking a balance between data utility and privacy. Under that scope, risk analysis and anonymization tools are required in order to increase awareness of the privacy risks, aid in regulatory compliance, and assist data processors with the anonymization process. Although there are a few tools reported in literature, they do not offer a wide range of options in terms of

the types of data that can be analyzed, the support of data multidimensionality, and visual exploration of the risk analysis results.

Aside from data privacy issues regarding the disclosures and exchanges of Big Data, there are also challenges over their meaningful analysis. Visual analytics is a research area that focuses on offering efficient and transparent methods of processing, visualizing, and analyzing large volumes of data so that analysts may better understand them and extract insights that could support data-driven decision making. In the literature, a variety of Visual analytics applications are available. Among the most common features of such applications is the ability to create dashboards in order to support Big Data exploration. Dashboards are a collection of data visualizations and filtering options designed to assist analysts and provide an interactive way for them to conduct their analysis. However, most of the currently available solutions fall short when it comes to dashboard-wide data exploration through drill-down or roll-up analysis. Data drill down refers to the process by which an analyst can shift from a grouping of data to a more detailed and granular group of data, whereas roll-up refers to investigating data in progressively less detailed levels. The applications offering this functionality only provide it in a limited fashion and for specific charts or graphs, without being able to support propagation of the drilling or rolling actions to the rest of the dashboard's visualizations.

Our proposed methodology for dealing with the aforementioned issues involves the design and development of a unified framework of applications aimed at the analysis, visualization, and exploration of big data while ensuring security and privacy. These applications provide the ability to analyze the risk of leaking personal data that may pass through a set of data, and also the ability to anonymize them. Furthermore, they facilitate the visualization and exploration of large datasets by combining previously owned datasets with those obtained from digital data marketplaces and displaying them through interactive dashboards. These dashboards can be adapted to the user's analysis framework requirements and provide data-drilling functionalities based on the type of data under analysis, thus allowing users to gain new insights that they could not have gained otherwise.

Σχεδίαση και ανάπτυξη ενιαίου πλαισίου ανωνυμοποίησης, ανάλυσης, οπτικοποίησης και εξερεύνησης μεγάλων δεδομένων τα οποία αντλούνται μέσω ψηφιακών αγορών

Περίληψη

Στη σημερινή εποχή που βασίζεται στα δεδομένα, η ανταλλαγή αυτών παίζει καθοριστικό ρόλο στην καθημερινότητά μας. Κάθε ψηφιακή συναλλαγή, από την πιο απλή έως την πιο περίπλοκη, απαιτεί ανταλλαγή δεδομένων μεταξύ των εμπλεκόμενων μερών. Από ιδιώτες και μικρές επιχειρήσεις έως μεγάλες εταιρείες, οργανισμούς και κυβερνήσεις όλοι αποθηκεύουν, επεξεργάζονται και ανταλλάσσουν δεδομένα. Αυτή η πραγματικότητα, με την πάροδο του χρόνου, οδήγησε στη συσσώρευση τεράστιων όγκων δεδομένων, γνωστά και ως Μεγάλα Δεδομένα. Με την έλευση των Μεγάλων Δεδομένων, έγινε φανερό ότι υπήρχαν πολλές ευκαιρίες αναφορικά με την ανάλυσή τους και ότι τα αποτελέσματα τέτοιων αναλύσεων θα μπορούσαν να παρέχουν στους επεξεργαστές των δεδομένων αυτών εξαιρετικά ωφέλιμες πληροφορίες (insights). Επίσης, μεγάλη βοήθεια στη βελτίωση των αποτελεσμάτων τέτοιων αναλύσεων μπορεί να προσφέρει ο εμπλουτισμός και η συσχέτιση των υπαρχόντων, ιδιόκτητων, συνόλων δεδομένων με σύνολα που προήλθαν από εξωτερικές πηγές. Η απόκτηση συνόλων δεδομένων μέσω τρίτων, ήταν κατά κανόνα μια διαδικασία που απαιτούσε την άμεση προσέγγιση συγκεκριμένων κατόχων δεδομένων. Ωστόσο, τα τελευταία χρόνια, με την εμφάνιση των ψηφιακών αγορών δεδομένων, αυτή η κατάσταση έχει αρχίσει να αλλάζει.

Στο πρόσφατο παρελθόν, οι ανταλλαγές δεδομένων πραγματοποιούνταν με ελάχιστη έως καθόλου μέριμνα για το απόρρητο ή την προστασία των προσωπικών δεδομένων. Πρόσφατες νομοθετικές εξελίξεις, όπως η νομοθεσία της Ευρωπαϊκής Ένωσης για την προστασία των προσωπικών δεδομένων ΓΚΠΔ, ώθησαν πολλούς παρόχους και καταναλωτές Μεγάλων Δεδομένων να αναζητήσουν λύσεις τόσο για την προστασία του απορρήτου όσο και για την αξιολόγηση των κινδύνων απορρήτου των συνόλων δεδομένων που αυτοί διαχειρίζονται. Μετά από αυτές τις εξελίξεις, οποιαδήποτε δημοσιοποίηση δεδομένων πρέπει να εφαρμόζει κάποια μορφή προσαρμογής των δεδομένων αυτών πριν από τη δημοσιοποίηση, ώστε να προστατεύει το απόρρητο των ευαίσθητων πληροφοριών των ατόμων. Η ανωνυμοποίηση δεδομένων είναι ένα παράδειγμα μιας τέτοιας διαδικασίας προσαρμογής και περιλαμβάνει την αφαίρεση ή τη

μετατροπή δεδομένων, με τρόπο που διατηρεί το απόρρητο, εξασφαλίζοντας έτσι ένα ορισμένο επίπεδο ανωνυμίας αυτών. Μία από τις πιο δύσκολες πτυχές οποιασδήποτε διαδικασίας ανωνυμοποίησης δεδομένων είναι η επίτευξη ισορροπίας μεταξύ της χρησιμότητας των δεδομένων και του απορρήτου. Υπο το πρίσμα αυτό, εργαλεία ανάλυσης κινδύνου και ανωνυμοποίησης είναι απαραίτητα προκειμένου να αυξηθεί η ευαισθητοποίηση σχετικά με τους κινδύνους απορρήτου καθώς και να βοηθηθούν οι επεξεργαστές των δεδομένων αυτών τόσο στη συμμόρφωσή τους με τους κανονισμούς όσο και στην ίδια την διαδικασία της ανωνυμοποίησης των δεδομένων αυτών. Αν και υπάρχουν κάποια εργαλεία στη βιβλιογραφία, αυτά δεν προσφέρουν ένα αρκετά ευρύ φάσμα επιλογών όσον αφορά τους τύπους δεδομένων που μπορούν να αναλύσουν, την υποστήριξη πολλαπλών διαστάσεων δεδομένων καθώς και την οπτική εξερεύνηση των αποτελεσμάτων των αναλύσεων που διενεργούνται.

Εκτός από τα ζητήματα απορρήτου των δεδομένων που υπάρχουν σχετικά με τις δημοσιοποιήσεις και συναλλαγές Μεγάλων Δεδομένων, υπάρχουν επίσης προκλήσεις σχετικά με την ουσιαστική ανάλυσή τους. Η Οπτική Ανάλυση (Visual Analytics) είναι ένας τομέας έρευνας που εστιάζει στην παροχή αποτελεσματικών και διαφανών μεθόδων επεξεργασίας, οπτικοποίησης και ανάλυσης μεγάλου όγκου δεδομένων, έτσι ώστε οι αναλυτές να μπορούν να τους κατανοήσουν καλύτερα και να εξάγουν πληροφορίες που θα μπορούσαν να υποστηρίξουν τη λήψη αποφάσεων. Στη βιβλιογραφία, υπάρχει αρκετή ποικιλία εφαρμογών Οπτικής Ανάλυσης. Ανάμεσα στα κοινά χαρακτηριστικά των εφαρμογών αυτών υπάρχει η δυνατότητα δημιουργίας ταμπλό (dashboards) για την υποστήριξη της εξερεύνησης Μεγάλων Δεδομένων. Τα ταμπλό (dashboards) είναι μια συλλογή οπτικοποιήσεων δεδομένων και επιλογών φιλτραρίσματος, σχεδιασμένα για να βοηθούν τους αναλυτές με την ανάλυση των δεδομένων παρέχοντας τους έναν διαδραστικό τρόπο για τη διεξαγωγή αυτής. Ωστόσο, οι περισσότερες από τις διαθέσιμες λύσεις, επί του παρόντος, υπολείπονται όσον αφορά τη εμβάθυνση (drill-down) και γενίκευση (roll-up) των δεδομένων που οπτικοποιούνται στο ταμπλό. Η εμβάθυνση στα δεδομένα αναφέρεται στη διαδικασία κατά την οποία ένας αναλυτής μπορεί να μεταβεί από μια ομαδοποίηση δεδομένων σε μια πιο λεπτομερή ομαδοποίηση, ενώ η γενίκευση αφορά στη διερεύνηση των δεδομένων σε σταδιακά λιγότερο λεπτομερές επίπεδο. Οι εφαρμογές που παρέχουν αυτήν τη λειτουργικότητα την παρέχουν με περιορισμένο τρόπο και μόνο σε

συγκεκριμένα γραφήματα ή γράφους, χωρίς να μπορούν να υποστηρίξουν τη διάδοση των ενεργειών της εμβάθυνσης ή γενίκευσης στις υπόλοιπες οπτικοποιήσεις του ταμπλό.

Η προτεινόμενη μεθοδολογία μας για την αντιμετώπιση των προαναφερθέντων ζητημάτων περιλαμβάνει τον σχεδιασμό και την ανάπτυξη ενός ενιαίου πλαισίου εφαρμογών που στοχεύουν στην ανάλυση, οπτικοποίηση και εξερεύνηση μεγάλων δεδομένων, διασφαλίζοντας παράλληλα την ασφάλεια και την ιδιωτικότητα. Αυτές οι εφαρμογές παρέχουν τη δυνατότητα ανάλυσης του κινδύνου διαρροής προσωπικών δεδομένων που μπορεί να διαρρεύσουν μέσα από ένα σύνολο δεδομένων, καθώς και τη δυνατότητα ανωνυμοποίησής τους. Επιπλέον, διευκολύνουν την απεικόνιση και την εξερεύνηση μεγάλων συνόλων δεδομένων συνδυάζοντας ιδιόκτητα σύνολα δεδομένων με σύνολα που αποκτήθηκαν από αγορές ψηφιακών δεδομένων και οπτικοποιώντας τα μέσω διαδραστικών ταμπλό. Τα ταμπλό αυτά μπορούν να προσαρμόζονται στις απαιτήσεις του πλαισίου ανάλυσης του χρήστη και να παρέχουν λειτουργίες εμβάθυνσης ή γενίκευσης των δεδομένων με βάση τον τύπο των δεδομένων υπό ανάλυση, επιτρέποντας έτσι στους χρήστες να εντοπίσουν νέα γνώση την οποία δεν κατείχαν πριν αναλύσουν τα συγκεκριμένα σύνολα δεδομένων.

Acknowledgements

I would like to thank my esteemed advisor – Professor Constantine Stephanidis for his invaluable supervision and support during our collaboration. I would also like to thank Dr. George Margetis for his unwavering support and extensive guidance throughout this thesis.

In addition, I'd like to express my gratitude to Dr. Stavroula Ntoa for her assistance in carrying out the evaluation of both the DaRAV and InfoDrill applications.

My gratitude also extends to my family and friends for their invaluable support and encouragement during this endeavor.

Dedicated to my parents and two brothers

Contents

| | |
|---|-----|
| Contents | i |
| List of Tables..... | v |
| List of Figures | vii |
| Introduction | 1 |
| Literature Review | 5 |
| 2.1 Data Privacy..... | 5 |
| 2.1.1 Anonymity of Data | 6 |
| 2.1.2 Privacy Models | 7 |
| 2.1.3 K-Anonymity..... | 8 |
| 2.1.4 Attacks on k-anonymity | 9 |
| 2.1.5 L-diversity and other k-anonymity extensions | 11 |
| 2.1.6 Differential Privacy..... | 12 |
| 2.1.7 Data Utilization and Risk Analysis | 13 |
| 2.1.8 Risk Analysis Tools and Limitations..... | 14 |
| 2.2 Visual Analytics..... | 16 |
| 2.2.1 Visual Analytics Affiliated Areas..... | 17 |
| Human Perception and Cognition..... | 17 |
| Analytical Reasoning | 19 |
| Data Representations, Transformations and Management | 20 |
| Data visualization, issues and approaches..... | 23 |
| 2.2.2 Visual Analytics Process and Knowledge Generation | 27 |
| 2.2.3 Data Visualization Applications and Limitations..... | 30 |
| Methodology..... | 33 |
| 3.1 Framework | 33 |
| 3.1.1 Framework Realization and Specifications | 34 |
| 3.2 Deanonimization Risk Analysis Application – DaRAV..... | 36 |

| | |
|--|----|
| 3.2.1 Application Description..... | 36 |
| Supported Datasets..... | 36 |
| Risk Analysis Modules | 37 |
| Anonymization Modules | 43 |
| 3.2.2 Implementation | 44 |
| Application Architecture | 44 |
| Technologies used..... | 46 |
| 3.2.3 Using the Application | 48 |
| User Login and Application Layout | 48 |
| Viewing the Available Datasets..... | 48 |
| Uploading a new Dataset | 50 |
| Editing a dataset..... | 50 |
| Starting a new Risk Analysis or Anonymization process..... | 51 |
| Viewing all initiated processes..... | 51 |
| Viewing the results of a process | 52 |
| Viewing the results of a Risk Analysis process..... | 52 |
| 3.3 Analytics and Insights Application – InfoDrill..... | 54 |
| 3.3.1 Application Description..... | 54 |
| About Datasets..... | 54 |
| Smart Dashboards..... | 55 |
| Data Drilling Visualizations..... | 56 |
| Complementary Visualizations | 58 |
| 3.3.2 Implementation | 60 |
| Application Architecture | 60 |
| Smart Dashboards and Dataset Correlation | 62 |
| Technologies used..... | 64 |
| 3.3.3 Using the Application | 66 |
| User Login and Application Layout | 66 |
| Uploading a new Dataset | 66 |

| | |
|--|-----|
| Viewing the Available Datasets..... | 67 |
| Combining Datasets | 68 |
| Editing Datasets | 69 |
| Create a new Dashboard..... | 70 |
| Dashboard Page | 72 |
| Create a new View | 73 |
| Filters..... | 74 |
| Evaluation..... | 77 |
| 4.1 Methodology..... | 77 |
| 4.1.1 Heuristic Evaluation of DaRAV Application..... | 77 |
| 4.1.2 User Testing Evaluation of InfoDrill Application | 79 |
| 4.2 DaRAV Heuristic Evaluation | 83 |
| 4.2.1 Procedure..... | 83 |
| 4.2.2 Results | 83 |
| 4.3 InfoDrill User Testing Evaluation..... | 86 |
| 4.3.1 Procedure | 86 |
| 4.3.2 Results | 87 |
| Task Success | 87 |
| Help Requests | 88 |
| Errors observed | 89 |
| Workload..... | 91 |
| User Experience | 92 |
| Debriefing findings | 93 |
| 4.4 Discussion..... | 95 |
| Conclusions and Future Work..... | 97 |
| 5.1 Conclusions | 97 |
| 5.2 Future work..... | 98 |
| References..... | 101 |

List of Tables

| | |
|---|----|
| Table 1 Scenario A: Dashboard Comprehension | 79 |
| Table 2 Scenario B: Dashboard Creation | 80 |
| Table 3 DaRAV Heuristic Evaluation Results | 83 |
| Table 4 Discovered Issues Description..... | 89 |

List of Figures

| | |
|--|----|
| Figure 1 Linking to re-identify data [5]..... | 7 |
| Figure 2 Example of applying 4-anonymity to a table of records [10] | 8 |
| Figure 3 Example of a domain (DGH_{M0}) and value (VGH_{M0}) generalization hierarchy of the marital status attribute [4]..... | 9 |
| Figure 4 Example of applying 3-diversity to a table of records [10]..... | 11 |
| Figure 5 Example of de-anonymization risk analysis performed in the ARX tool [15, 16] | 14 |
| Figure 6 Examples of the “The Law of Continuity” from the Gestalt Laws (left image) where the dots can be perceived as continuous lines and of Pre-attentive processing (right image) where the “pop out” effect can be observed. | 18 |
| Figure 7 The analytical reasoning process [22]..... | 19 |
| Figure 8 Examples of visualization techniques, a histogram diagram on the left and a radar chart on the right | 23 |
| Figure 9 The Visual Analytics Process [18]..... | 27 |
| Figure 10 Visual Analytics process and the Knowledge generation model [30]..... | 28 |
| Figure 11 Tableau Desktop..... | 31 |
| Figure 12 Proposed Framework..... | 33 |
| Figure 13 Results Page – The results of a risk analysis process with k-anonymity ($k=2$)..... | 38 |
| Figure 14 Results Page – The results of a risk analysis process with l-diversity ($l=2$)..... | 39 |
| Figure 15 Results Page – The results of a risk analysis process of location data for $w=1$, $r=200m$, $t=6$ hours | 39 |
| Figure 16 Results Page – The results of a risk analysis process of textual data | 40 |
| Figure 17 Results Page – The results of a risk analysis process of transactions data ($w=1$, $a=1000$, $t=1$ month) | 41 |
| Figure 18 Results Page – The results of a risk analysis process on aggregated data..... | 42 |
| Figure 19 DaRAV Application Architecture diagram..... | 44 |
| Figure 20 Datasets Page – card view | 49 |
| Figure 21 Upload Dataset Page..... | 49 |
| Figure 22 Dataset Info Page | 50 |
| Figure 23 Risk Analysis Page | 51 |
| Figure 24 Processing Queue Page..... | 52 |
| Figure 25 De-anonymization Risk Analysis Page..... | 53 |
| Figure 26 Smart Dashboard Example..... | 55 |
| Figure 27 Data Drilling Visualizations, a geographical map (left image) and a histogram with a date brush (right image) | 57 |

| | |
|--|----|
| Figure 28 Examples of Complementary Visualizations, a radar chart (top-left), a violin chart (top-right), a donut chart (bottom-left) and a bar chart(bottom-right)..... | 58 |
| Figure 29 InfoDrill Application Architecture diagram | 60 |
| Figure 30 Smart Dashboard implementation diagram | 62 |
| Figure 31 Upload Dataset Page..... | 67 |
| Figure 32 Datasets Page – all datasets tab – card view | 68 |
| Figure 33 Datasets Page – Combining Datasets..... | 68 |
| Figure 34 Datasets Page – Combined Datasets | 69 |
| Figure 35 Dataset Info Page | 70 |
| Figure 36 Create Dashboard page..... | 70 |
| Figure 37 Dashboard page | 72 |
| Figure 38 Create View page | 73 |
| Figure 39 Dashboard page – Filters | 74 |
| Figure 40 Task Success for scenario A (left) and B (right) | 87 |
| Figure 41 Percentage of users who completed the task without any assistance..... | 88 |
| Figure 42 Frequency of occurrence for each identified error during users' scenario executions | 89 |
| Figure 43 Mental Workload of InfoDrill Application for scenarios A & B versus Desktop applications (left) and Office work in general (right) [48] | 91 |
| Figure 44 UMUX-Lite Results (with error bars representing 95% confidence intervals)..... | 92 |
| Figure 45 UMUX-Lite Analysis of Responses..... | 92 |

Chapter 1

Introduction

With the arrival of the digital age in the second half of the 20th century, industries, institutions, and governments all over the world began to adjust many of their long-standing processes in order to adapt to this new setting. They gradually transitioned from physical and even handwritten archives to digitally stored archives while automating many of their manual and time-consuming procedures. In later years, with the advent of the internet, a similar transition occurred. Transactions and services that were previously conducted in person became digital and were delivered through websites and applications. As a result of these changes, the volumes of data collected and stored by the parties involved began to increase dramatically. Because of this phenomenon, as well as the peculiarities and characteristics surrounding these data, the scientific community started to refer to these massive amounts of data as Big Data.

With the emergence of Big Data, it became clear that there were numerous opportunities to be attained from their analysis. Many companies and organizations began to derive value from Big Data by applying the findings of their analyses to improve various aspects of their operations, such as decision making and customer experience while governments started to improve their efficiency and reduce operating costs. Under this scope, the correlation of data from different sectors was proven to be highly profitable since it could help with the indication of new aspects to long-standing problems as well as the identification of profitable groupings that could then be targeted through agile marketing activities.

Correlation or enrichment of existing internal datasets with datasets obtained from external sources used to rely on the direct approach and negotiation with specific data owners. This situation began to change in recent years, with the emergence of digital data markets. These

platforms enable their users to become either data providers or consumers on a mostly cloud-based market. There, the users can offer or seek datasets that meet their analysis needs, and make transactions in a more formal and direct manner, thus speeding up a process that could otherwise take considerable time. Because of the ease with which data could be exchanged and analyzed, many were the governments and institutions who were concerned about the security and privacy of entities or individuals who were directly or indirectly connected to these data.

Until recently, data owners conducted exchanges with little to no regard for privacy or the protection of personal data. This led to many legislative initiatives around the world, with one of the most prominent ones being the European Union's data protection regulation, GDPR [1]. Under this regulation, the exchange or transfer of personal data fell under many restrictions and within narrow exceptions.

As a result of these developments, any data disclosure within the EU must use some form of data sanitization prior to release in order to protect individuals' sensitive information and remain GDPR compliant. Data anonymization is one example of a sanitization process that GDPR recognizes. Anonymization of data procedures are typically based on the deduction or generalization of data in a privacy-preserving manner in order to achieve a certain level of anonymity.

In the literature there are many offerings regarding procedures for making data anonymous. These procedures are typically described by what are known as privacy models. Privacy models can differ on their approach to achieving anonymity and are usually intended for specific privacy threats. K -anonymity, l -diversity, t -closeness as well as differential privacy, are examples of such models. However, GDPR does not specify the means of making data anonymous, i.e., which privacy models should be used and in which cases. Furthermore, the regulation implies that the data owner needs to become aware of the risks of potential privacy breaches in the dataset [2]. This prompted many data owners to seek solutions for both protecting individuals' privacy and assessing the privacy risks of the datasets under their management. Although there are a few tools reported in literature, they do not offer a wide range of options in terms of the types of

data that can be analyzed, the support of data multidimensionality, and visual exploration of the risk analysis results.

Another crucial aspect of Big Data is their meaningful analysis and the potential value and insights that can be derived from it. One of the research fields studying the topic of Big Data is that of Visual analytics. Visual analytics focus on the processing, visualization, and analysis of large amounts of data. It is a multidisciplinary field that encompasses various disciplines such as human perception and cognition, analytical reasoning, and interactive data visualization as well as more technical ones such as data representations and management. All of these disciplines are being employed through the Visual analytics process in order to aid in the knowledge generation. The disciplines derived from this process are directly applied to applications aimed at providing analytics and insights through Big Data analysis. Such applications exist in a variety of types. They typically provide their functionalities through the use of dashboards, from which users can create custom visualizations for attributes of the data they want to analyze as well as apply filters.

However, when it comes to dashboard-wide propagation of data drill-down or roll-up actions, the majority of currently available solutions falls short. Data drill down is the method by which an analyst may move from a grouping of data to a more specific and granular group of data, whereas data roll-up is the process of exploring data at successively lower levels of detail [3]. Currently, this functionality is typically provided by single visualizations such as maps and charts with users having to perform extensive customizations in order for the drill down or roll up operations they perform to be propagated to the rest of the dashboard's visualizations.

The motivation for this thesis stems from the importance of privacy protection and the need for data owners and controllers to become aware of the potential privacy risks that their datasets may pose. This is especially crucial when it comes to datasets that will be exchanged on digital marketplaces. At the same time, we want to emphasize the necessity of correlating and enhancing existing datasets with datasets obtained from such data markets, as well as the potential value and insights that can be extracted from analyzing them.

This thesis discusses a methodology for dealing with the aforementioned issues by providing a unified framework aimed at the analysis, visualization, and exploration of big data while ensuring

security and privacy. Specifically, the proposed framework realizes two applications, called DaRAV and InfoDrill. Through the first application, we aim at providing a solution that can facilitate de-anonymization risk analysis modules and anonymization functionalities so as to help data owners analyze the risk of leaking personal data that may pass through a set of data, and also offer them the ability to anonymize them. With the second application we aim at providing a solution for the visualization and exploration of large datasets by combining previously owned datasets with those obtained from digital data marketplaces and displaying them through smart interactive dashboards. These dashboards will be able to adapt to the user's analysis framework requirements and provide data drill down and roll up functionalities based on the type of data under analysis, thus allowing users to gain new insights that they could not have gained without analyzing those datasets.

The rest of the thesis is organized as follows:

- **Chapter 2 – Literature Review:** in this chapter we will review related work regarding data privacy and visual analytics while shading a light on their most known practices.
- **Chapter 3 – Methodology:** introduction and discussion on the framework, the applications that the framework comprises and the functionalities they provide.
- **Chapter 4 – Evaluation:** evaluation of the framework with users and experts from the field of data analysis as well as our findings.
- **Chapter 5 – Conclusions and Future Work:** thesis conclusion, reviewing the contributions of our work and future work.

Chapter 2

Literature Review

This chapter focuses on two important aspects regarding the framework we propose. The first aspect concerns data privacy and what are the options available for data controllers in order to anonymize the data under their management. Also, we discuss the issue of data utilization versus the de-anonymization risks and lastly the tools available for conducting risk analysis and anonymization procedures as well as their limitations. The second aspect addresses the topic of visual analytics and how this research field is approaching the issue of Big Data analysis as well as the processes involved, the applications offered and their limitations.

2.1 Data Privacy

In today's data driven world we, the users, are provided with a wide variety of digital products and services that we access daily through our computers, handheld devices and even wearables. Those applications and services are being offered from various parties. From leading technology companies and organizations to governments and even local businesses, all offer a comprehensive variety of software that aims at helping our everyday lives.

In order for these parties to provide their digital products and services, they usually have to keep records of the individuals who are using them. Records that usually contain personal or sensitive information, information that should and has to be protected.

Many are the occasions though, that such parties need to disclose the records that they possess to third parties, due to a diversity of different needs that pertain to dependences between the parties involved. For example, a financial institution or a bank might want to improve its credit score models. So, for the purpose of gaining previously unknown insights from the data it

possesses, it discloses the financial information of its customers to an analytics firm capable of analyzing them. Another example might regard data exchange for research purposes, e.g. a hospital could release clinical data of patients with a particular illness to clinical laboratories in order to assist them on their effort to address an epidemic.

In order for such a disclosure of records, that include private or sensitive information, to commence, some form of sanitization to the data needs to take place, in advance. The techniques used for data privacy preservation today are many. This work will mainly focus on Data Anonymization, as a technique of achieving the privacy preservation of data.

2.1.1 Anonymity of Data

Data Anonymization is generally considered as the process in which the data about to be disclosed are either deducted or transformed (usually through generalization) in a privacy preserving manner, thus achieving a certain, pre-decided, level of anonymity.

Some of the first, and also naïve, approaches regarding the anonymization of data, considered records as anonymized only when the explicit information that could uniquely identify individuals, like their name, social security number or telephone, were removed. Those approaches, though, were quickly proven to be inadequate to protect the individuals' personal information since they did not take into account Quasi-Identifiers (QIs). QIs are the attributes of a record that do not directly identify an individual, but, when combined, could serve as a unique identifier [4]. Thus, those approaches were susceptible to many re-identification attacks and practically provided a false sense of security.

One of the most notable de-anonymization incidents was the re-identification of personal data of state employees in the state of Massachusetts, U.S. by L. Sweeney in 1997 [4]. At that time, the Group Insurance Commission (GIC) in Massachusetts had collected patient-specific data needed for purchasing health insurance for state employees. These data, which GIC considered to be anonymous (since they didn't include explicit identifiers like names or telephone numbers), were given to researchers as well as sold to the industry. Those records, besides containing the health data of state employees, also contained their ZIP code, birth date and sex information.

Those attributes were enough though, as L. Sweeney showed, in order to link those records with the records of a local voting registration list.

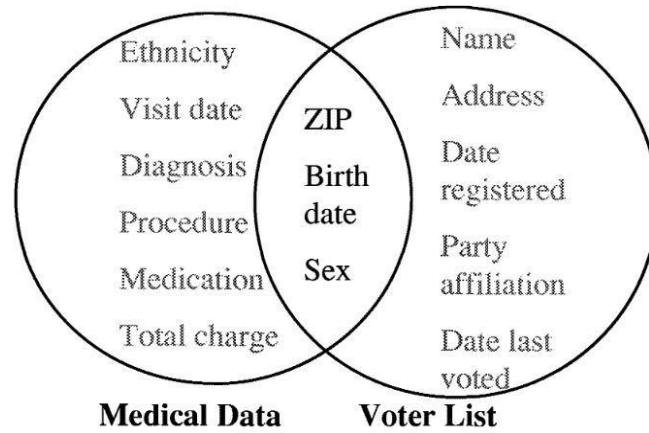


Figure 1 Linking to re-identify data [5]

Hence, by using the ZIP code, birth date and gender as common attributes between the two datasets, the diagnosis, procedures, and medications records of the GIC dataset could be linked to particular individuals from the voters list. This highlighted the unavoidable exposure of sensitive information of the state employees, with one of them even being the then Governor of Massachusetts. This also meant that the once considered anonymized records of the GIC dataset were proven to provide only a false sense of security and that the removal of only the individual's explicit identifiers was not enough for achieving true anonymity. To aggravate the situation further, in another study, Sweeney [5] showed that 87% of the U.S. population was uniquely identifiable by their ZIP code, gender and date of birth information.

Other famous de-anonymization incidents were also those of the AOL search data leakage in 2006, where individuals were identified through their AOL site searches [6] and the identification of individual Netflix users from the Netflix Prize Dataset by Narayanan and Shmatikov in 2007 [7].

2.1.2 Privacy Models

To preserve the privacy of individuals while disclosing records so as to prevent incidents as the ones mentioned previously, some form of anonymization needs to be employed to these records prior to their release. In order for the criteria of an anonymization procedure to be precise and

for the technique itself to be reproducible, the notion of the Privacy Model was introduced. A Privacy Model can be considered as an Ex-Ante statement of the privacy guarantees that can be obtained from a set of records by applying to it one or several anonymization techniques [8].

In the literature there is a great variety of Privacy Models which are applied depending on the use case and the type of data. Such privacy models are k-anonymity and its extensions (*l*-diversity and t-closeness), as well as the alternative paradigm of differential privacy.

2.1.3 K-Anonymity

One of the first privacy models created to address the re-identification problems caused by linking attacks was k-anonymity. It was proposed by Samarati and Sweeney [9, 4], who defined it as a model that requires that every record in a table of data is indistinguishable from at least k-1 other records in terms of its QI attributes, thus resulting in equivalence classes of at least size k.

| | Non-Sensitive | | | Sensitive | | Non-Sensitive | | | Sensitive |
|----|---------------|-----|-------------|-----------------|----|---------------|------|-------------|-----------------|
| | Zip Code | Age | Nationality | Condition | | Zip Code | Age | Nationality | Condition |
| 1 | 13053 | 28 | Russian | Heart Disease | 1 | 130** | < 30 | * | Heart Disease |
| 2 | 13068 | 29 | American | Heart Disease | 2 | 130** | < 30 | * | Heart Disease |
| 3 | 13068 | 21 | Japanese | Viral Infection | 3 | 130** | < 30 | * | Viral Infection |
| 4 | 13053 | 23 | American | Viral Infection | 4 | 130** | < 30 | * | Viral Infection |
| 5 | 14853 | 50 | Indian | Cancer | 5 | 1485* | ≥ 40 | * | Cancer |
| 6 | 14853 | 55 | Russian | Heart Disease | 6 | 1485* | ≥ 40 | * | Heart Disease |
| 7 | 14850 | 47 | American | Viral Infection | 7 | 1485* | ≥ 40 | * | Viral Infection |
| 8 | 14850 | 49 | American | Viral Infection | 8 | 1485* | ≥ 40 | * | Viral Infection |
| 9 | 13053 | 31 | American | Cancer | 9 | 130** | 3* | * | Cancer |
| 10 | 13053 | 37 | Indian | Cancer | 10 | 130** | 3* | * | Cancer |
| 11 | 13068 | 36 | Japanese | Cancer | 11 | 130** | 3* | * | Cancer |
| 12 | 13068 | 35 | American | Cancer | 12 | 130** | 3* | * | Cancer |

Figure 2 Example of applying 4-anonymity to a table of records [10]

To enforce k-Anonymity, a combination of data generalization and suppression is employed to meet the model's requirements while still keeping the data useful for the recipients.

Data generalization is a technique that employs generalization relationships based on domain and value generalization hierarchies (see Figure 3) of Quasi-Identifying attributes in order to develop generalization strategies that allow for the substitution of QI attributes values with less specific ones, resulting in equivalence classes within the dataset [9, 4]. While suppression of data

is the technique which removes selected data values from the table, as needed, in order to achieve the privacy models' constraints [9].

Typically, the aforementioned strategies are used collaboratively in order to achieve the anonymity requirements while maintaining data integrity. In spite of that, some argue that data suppression should be regarded as a complementary approach rather than a primary one due to its implications on data utility [9].

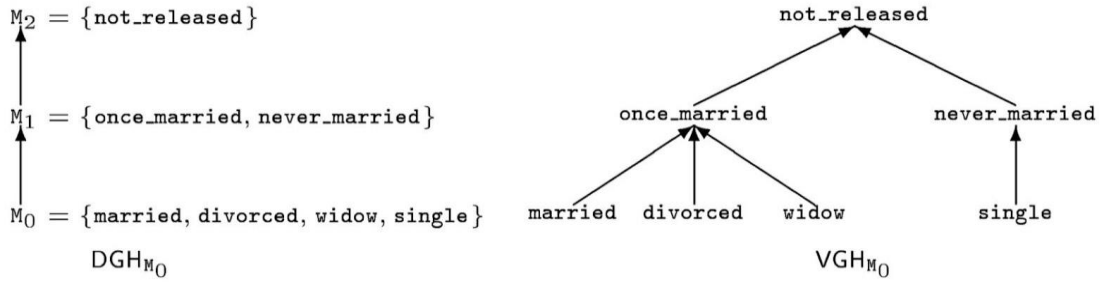


Figure 3 Example of a domain (DGH_{M_0}) and value (VGH_{M_0}) generalization hierarchy of the marital status attribute [4]

As shown in Figure 3, k-anonymity is capable of protecting against identity disclosures, since it ensures the creation of equivalence classes of at least size k. In this way, an attacker cannot identify an individual precisely. This, combined with the approach's relative conceptual simplicity, led to widespread adoption over time. It also led to many works in the literature that improved the original algorithm in terms of complexity and data types that it can handle.

2.1.4 Attacks on k-anonymity

However, as shown by Machanavajjhala, Ashwin, et al. [10] in 2007, k-anonymity's privacy protection of a dataset's records is only limited to identity disclosures since it provides no protection against attribute disclosures. This problem results in a dataset's k-anonymized records being exposed to a variety of attacks.

One of those attacks is the Homogeneity Attack [10]. In this type of attack, an attacker with access to a k-anonymized dataset can deduce the values of sensitive attributes by taking advantage of

the scenario when those sensitive attributes' values are all the same within an equivalence class of records. For example, let us consider two individuals Jane and John who are neighbors and live in the 13045 zip code. One day, John is admitted to the hospital. Jane notices and wants to learn more about John's illness. From the 4-anonymized table, as shown in Figure 2, Jane can deduce that one of the last four records (9 - 12) could be John's since Jane is aware that John is roughly 35 years old and that they both live in the 13045 zip code. Since the condition attribute for all those records is the same, Jane can confirm that the illness that John has is cancer.

Another attack that was identified was the Background Knowledge Attack [10]. By correlating QI attributes with prior knowledge about the individuals, an attacker with access to the k-anonymized dataset can infer the values of sensitive attributes by narrowing down the possible values of the attributes. So, continuing with the previous example, let us consider that Jane has a friend called Jack, who is a 28-year-old male that currently lives in the zip code 13056. Jack was also admitted to the same hospital as John. From the 4-anonymized table, as shown in Figure 2, Jane can, again, deduce that one of the first four records (1 - 4) could be Jack's since Jane is aware that Jack is 28 years old and that he lives in postal code 13056. In this case, though, Jack could be suffering from either heart disease or a viral infection. Jane is aware that Jack's father and grandfather have suffered from heart disease, and as a result, she concludes that Jack is very likely to suffer from heart disease as well.

These attacks highlighted the privacy difficulties that k-anonymity cannot address, which is why Machanavajjhala, Ashwin, et al. [10] introduced *l*-diversity, a new and more protective privacy model as an extension of the k-anonymity model.

2.1.5 L-diversity and other k-anonymity extensions

The purpose of the l -diversity model is to attempt to address the privacy issues that k -anonymity has against attribute disclosure. Specifically, this privacy model requires a minimum level of

| Non-Sensitive | | | | | Non-Sensitive | | | | |
|---------------|----------|-----|-------------|-----------------|---------------|----------|-----------|-------------|-----------------|
| | Zip Code | Age | Nationality | Condition | | Zip Code | Age | Nationality | Condition |
| 1 | 13053 | 28 | Russian | Heart Disease | 1 | 1305* | ≤ 40 | * | Heart Disease |
| 2 | 13068 | 29 | American | Heart Disease | 4 | 1305* | ≤ 40 | * | Viral Infection |
| 3 | 13068 | 21 | Japanese | Viral Infection | 9 | 1305* | ≤ 40 | * | Cancer |
| 4 | 13053 | 23 | American | Viral Infection | 10 | 1305* | ≤ 40 | * | Cancer |
| 5 | 14853 | 50 | Indian | Cancer | 5 | 1485* | > 40 | * | Cancer |
| 6 | 14853 | 55 | Russian | Heart Disease | 6 | 1485* | > 40 | * | Heart Disease |
| 7 | 14850 | 47 | American | Viral Infection | 7 | 1485* | > 40 | * | Viral Infection |
| 8 | 14850 | 49 | American | Viral Infection | 8 | 1485* | > 40 | * | Viral Infection |
| 9 | 13053 | 31 | American | Cancer | 2 | 1306* | ≤ 40 | * | Heart Disease |
| 10 | 13053 | 37 | Indian | Cancer | 3 | 1306* | ≤ 40 | * | Viral Infection |
| 11 | 13068 | 36 | Japanese | Cancer | 11 | 1306* | ≤ 40 | * | Cancer |
| 12 | 13068 | 35 | American | Cancer | 12 | 1306* | ≤ 40 | * | Cancer |

Figure 4 Example of applying 3-diversity to a table of records [10]

diversity for the sensitive attribute in each of the k -anonymous groups of records (i.e., equivalence classes) [10]. In this way, a dataset can maintain its anonymous integrity even when the data publisher does not know what kind of knowledge is possessed by the adversary / attacker.

In greater detail, let's consider a q^* -block to be a set of tuples in the table records whose non-sensitive values generalize to a value q^* . Then, a q^* -block is considered to be compliant to l -diversity if it has at least l ($l \geq 2$) different sensitive values. If this is true, the q^* -block is regarded as “well-represented” [10] by l sensitive values, and if this is true for the rest of the dataset's q^* -blocks, the dataset is deemed to be l -diverse. The results of such a procedure can be seen in Figure 4, where the new 3-diverse table has at least 3 different values of the sensitive attribute “Condition”.

By enforcing a minimum amount of variability for the sensitive attributes in each equivalence class, l -diversity attempts to reduce the risk of attribute disclosure. However, as the literature has shown, even the l -diversity model is unable to prevent some attacks with some even arguing that it is difficult and sometimes even unnecessary to achieve.

N. Li et al. [11] argued that l -diversity, while seeking to increase the number of distinct values per equivalence class, doesn't take into consideration the fact that the distribution of an attribute might be skewed. For example, a dataset might have records that contain a sensitive attribute that indicates the presence or absence of a given disease. Assuming that 99% of the individuals are negative, in order to enforce 2-diversity, the percentage of positive individuals should increase to 50%, thus dramatically increasing the probability of an individual being positive on the resulting equivalence class.

Another issue raised by N. Li et al. [11] was that, while l -diversity ensures that each equivalence class contains different values of a sensitive attribute, it does not defend against different but semantically similar sensitive attribute values. For example, if the domain of a sensitive attribute's values contains the values "myocardial infarction", "coronary thrombosis" and "cardiac arrest" it is simple for an adversary to understand that an individual has a heart related problem.

After identifying those issues, N. Li et al. [11] offered a new Privacy Model that acted as an extension to l -diversity called t -closeness. They defined it as the model that ensures each sensitive attribute of an equivalence class has a distribution distance of no more than a threshold t with the distribution of the same attribute in the whole dataset. In this way they were able to take into account the distribution of sensitive attribute values and thus deal with the issues that were raised.

2.1.6 Differential Privacy

Differential privacy belongs to a different group of privacy models where the model doesn't focus on anonymizing the whole dataset nor is affected by the records and attributes of the dataset or the existence of other datasets that could be used for linkage. This type of model focuses on the Uninformative Principle [10] according to which the published dataset should not provide new information to an adversary, thus keeping the differences between their prior and posterior beliefs minimal. The Differential privacy model is described as an interactive privacy mechanism for queries submitted to databases that contain individual records.

Dwork [12] in 2006, proposed a mathematical definition of Differential Privacy called ϵ -differential privacy. Specifically, let us consider a randomized function “K” that takes a dataset as input and ϵ being a real number. The function is said to be fulfilling the promises of Differential Privacy if for all datasets D1 and D2 that differ on at most one record, and for all $S \subseteq \text{Range}(K)$, where $\text{Range}(K)$ is the set of possible outputs of the function, holds true that:

$$\text{Probability}[K(D1) \in S] \leq \exp(\epsilon) \times \text{Probability}[K(D2) \in S]$$

The main premise of this model is that the probability of an individual's identity being disclosed should not be significantly increased because they participated in a statistical database. As a result, the addition or deletion of a single record from the database has no effect on the outcome of any analysis and the record of an individual does not alter the results of the anonymization algorithm or the query performed on the database. In this way, an adversary cannot know if an individual's information was used for deriving the results published.

2.1.7 Data Utilization and Risk Analysis

Conforming to privacy models comes at the expense of a loss in a dataset's usefulness. The more private a dataset is, the more distorted it gets, and therefore the less valuable it is for the recipient. This is the reason why all the privacy models offer parameters to the owner of the dataset in order to choose the level of security they desire. For example, in the k-anonymity Privacy Model the user can specify the parameter “k” to the value they desire, with lower values meaning smaller equivalence classes and thus less secure but more usable datasets. In contrast, higher “k” values mean larger equivalence classes and thus more secure but less usable datasets since the algorithm has to increase the use of generalization for the QI attributes’ values and even reach the point of having to suppress them.

This leads to one of the great challenges in Data Anonymization, which is to strike a balance between Data Privacy and Data Utilization. This decision usually regards the data owner and their judgement on choosing the correct privacy models for the data as well as configuring their parameters to fit the privacy requirements needed. In the literature there is only a limited work in proposing models on deciding the privacy-utility trade-offs such as [13, 14] since many

CHAPTER 2. LITERATURE REVIEW

parameters like the type of data and value to be extracted from the dataset at hand need to be taken into account when deciding [2].

In order for a data owner to make an informed decision on those issues, they need to be able to realize the risk of a potential attack [2]. Usually the de-anonymization attack risks of a dataset are dependent upon the effort, time, costs, and experience of the attacker, with those who offer more having higher chances of actually achieving a successful one. So, the owner of the data should both gather external information that could be leveraged against the privacy of the released data as well as use tools that could indicate the de-anonymization risks in their datasets.

2.1.8 Risk Analysis Tools and Limitations

For a range of factors, using Risk Analysis Tools before disclosing data could be beneficial. In general, these tools can detect and highlight de-anonymization threats in datasets, as well as the extent to which they are de-anonymizable. They can also assist with the anonymization process by helping determine the necessary anonymization parameters since they can indicate the distortion required for a dataset to conform with a specific privacy model.

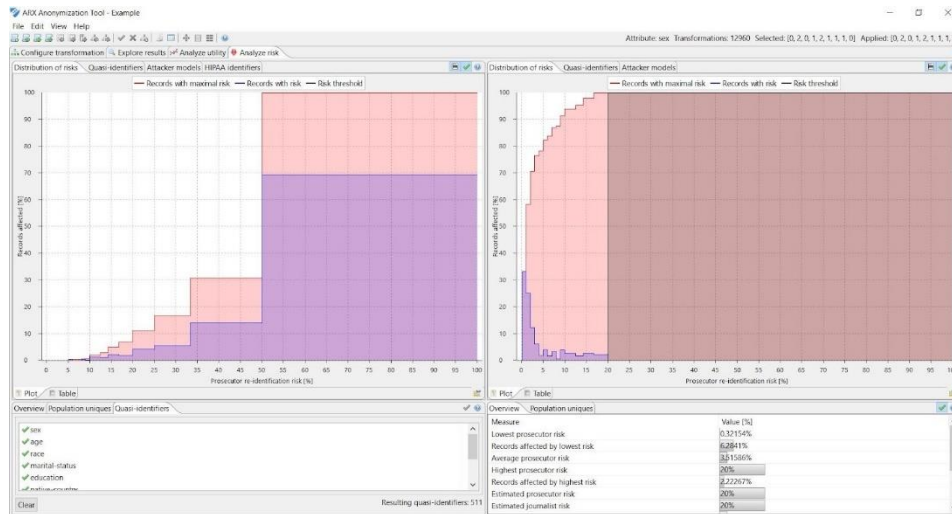


Figure 5 Example of de-anonymization risk analysis performed in the ARX tool [15, 16]

In the literature there are only a few tools that can provide those functionalities. Two of them are the ARX Data Anonymization Tool [15, 16] and X²R² [17]. The ARX tool is considered to be the current state-of-the-art tool for conducting anonymization and de-anonymization risk analysis procedures and it offers a range of privacy models for anonymization like k-anonymity, *l*-diversity

and differential privacy among others. The X^2R^2 (Explainable Explorative Re-Identification Risk) tool [17] is a data anonymization tool that aims at the non-experts. It provides risk analysis on the data imported and offers ways of addressing those risks by recommending transformations on the data and it gives real time feedback on the privacy risk reduction and data quality.

However, the main limitation of these tools is that they only deal with the risk analysis of tabular data, in which one record corresponds to one individual, and do not offer analyses that can deal with complex, high-dimensional and of specific nature data such as spatiotemporal or transaction data. This has implications for the owners of such datasets since they are left with limited, generic, options when it comes to evaluating their datasets against privacy risks and attacks.

2.2 Visual Analytics

With the sharp increase of digital human activity over the past decades, massive sets of data have been collected and stored by various parties. These volumes of data can be dynamic, ambiguous, incomplete and even conflicting. Over the course of time, these data also tend to become larger and larger in size and thus start to pose significant challenges to their owners in terms of their optimal utilization. Such challenges are primarily located in the data management and analysis areas where the conventional tools used for storage and analysis are becoming increasingly cumbersome to use for traditional users like analysts and business people.

When it comes to data analysis, such issues typically arise as a result of the information overload that comes with dealing with massive amounts of data. Information overload [18] can be considered as the issue where analysts, or data reviewers in general, are getting “lost” in the data that are in the process of analyzing. This can result from data that are either unrelated to the current task in hand or are processed/presented in an ineffective manner. The effects of information overload, besides the obvious stress that is inflicted upon the analysts themselves, can also be the cause of wasted opportunities since, in most businesses, success is heavily dependent on the availability of the right information at the right time.

One of the options that could mitigate these problems seems to be the use of more powerful tools for conducting data analyses. The issue with this option is that these tools tend to be oriented to automated data analyses. Such kinds of analyses come with their own set of issues though. Besides the problem of whether the user is able to understand their results and derive insights from them, there is also the issue of who has the responsibility of the derived results and the consequent decisions taken based on them. Specifically, it is important for any data owner, in the case where a false decision has emerged, to be able to examine the processes that were responsible for deriving the results that this decision was based upon. In this way, processes can be changed or improved thus preventing any wrongful decisions in the future. In order to achieve that though, the processes followed for the creation of the results have to be transparent and clear to all the stakeholders involved and not only known to the creators of the automated system or tool.

Visual Analytics focuses on providing effective and transparent ways of processing and analyzing those large volumes of data in order to enable the reviewers to understand them and derive insights more effectively. Specifically, Visual Analytics is a multidisciplinary field that aims at tackling the aforementioned issues by combining analytical processes with interactive visual interfaces in order to support data-driven decision making and analytical reasoning. It is a blend of the analytic processes and algorithms provided by modern computer systems and the knowledge of human analysts that can lead to detecting the expected results as well as breakthroughs and discoveries of hidden - or hiding in plain sight – insights [19].

2.2.1 Visual Analytics Affiliated Areas

Visual Analytics is a multidisciplinary field that, at its core, encompasses the fields of Information and Scientific Visualization, Data Management and Data Analysis. These fields are the pillars of Visual Analytics since the tools and techniques developed on this scope have been based on the research findings affiliated with these areas. But, on a broader scope, when someone conducts a research on the Visual analytics field, one has to also consider other relative fields and principals like analytical reasoning and human computer interaction as well. The following sections intend to highlight the important aspects of the aforementioned areas as well as their challenges and possible solutions.

Human Perception and Cognition

Visual Analytics is based on human perception and cognition. All of the field's applications will succeed or fail based on whether or not they enable the user to perceive and understand the information being delivered. Perception, and Visual Perception in particular, as well as Cognition are concepts and theories that tend to venture out of the scientific scope of Computer Science and more into the area of Psychology.

In Psychology, Perception is defined as the process of attaining awareness or understanding sensory information [20]. Visual Perception, as a subpart of Perception, is the ability to interpret the surrounding environment through the sense of vision. Psychological research on Perception

of visual information and human cognition focuses mainly on the ways in which individuals understand what they see in their environment.

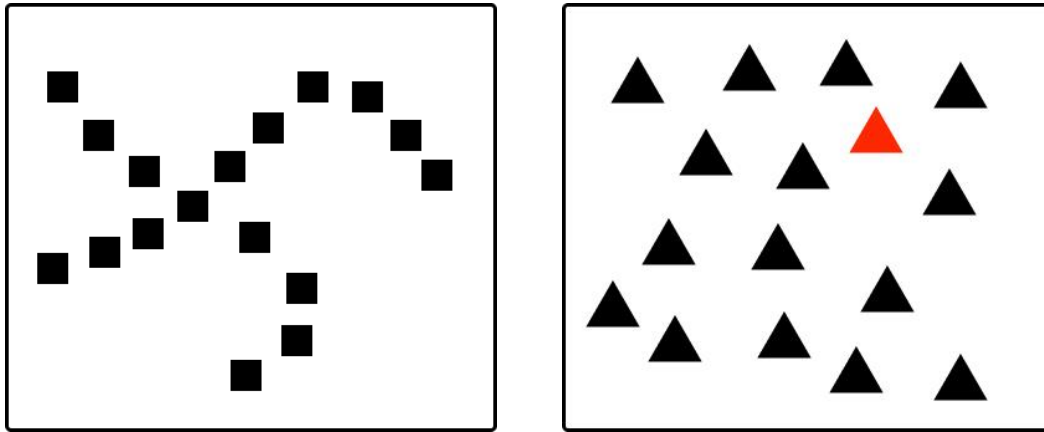


Figure 6 Examples of the “The Law of Continuity” from the Gestalt Laws (left image) where the dots can be perceived as continuous lines and of Pre-attentive processing (right image) where the “pop out” effect can be observed.

Gestalt (German word for 'shape' or 'form') psychology is one of the most influential schools of thought in Psychology with regards to perception. Gestalt psychologists emphasized that individuals can perceive entire patterns or configurations and not just individual components or parts of visual components. One of the theories of that school is The Gestalt Laws of Organization [20]. The organization of perceptual scenes, according to this theory, is based on rules that can determine how an individual groups elements into patterns, some of which are proximity, similarity, closure, symmetry, and continuity (see Figure 6, left image). These rules are used by designers of visual representations in order to efficiently guide the attention of the users through the graphical user interface as well as help them focus on elements of importance.

Another psychological theory related to visual perception is that of Pre-attentive Processing [21]. In this theory, information is being accumulated and pre-attentively processed subconsciously with information of high prominence being selected for further and more complete analysis from the conscious processing part of the brain. In the context of Visual analytics, this theory can explain the fact that some elements of visual displays “pop out” and can be easily distinguished

from others (see Figure 6). Such effects can greatly assist the process of information visualization since they can support visual search considerably [18].

Perception and visual perception in particular is a very broad field that is pertained by many theories and disciplines that can help us understand human perception and cognition as well as greatly assist in the implementation of effective visual information systems. Research on this field has led to the development of principles and methods that have greatly assisted to the design of perception-driven interactions between the users and the system, thus enabling the exploration of large information spaces to be conducted much more effectively.

Analytical Reasoning

Analytical reasoning is the ability of an individual to notice patterns within a set of facts or rules and utilize these observations in order to derive insights that should hold true. As a science, analytical reasoning can provide a framework on which Visual Analytics applications and procedures can be developed with specific focus on the human reasoning aspect.

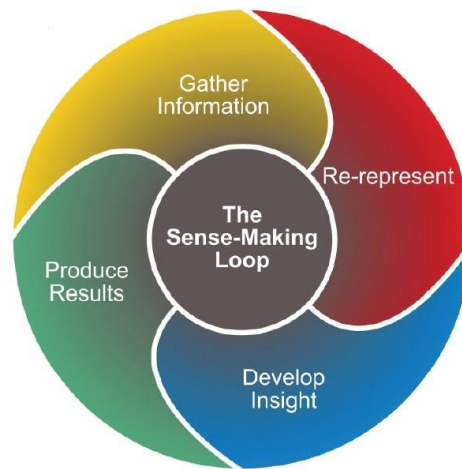


Figure 7 The analytical reasoning process [22]

As a process, analytical reasoning is pivotal to the field of Visual analytics since one of the main goals of the technologies developed under this scope is to maximize the ability of an individual to perceive, understand and finally reach to discoveries and insights from the large volumes of diverse and dynamic data at hand. To that extent, high-quality human judgment with a limited

investment of the analysts' time should be facilitated from such technologies in order for the analysts to be able to infer insights that can effectively support situation assessment, planning, and decision making [22].

A scientific field closely related to analytical reasoning is that of sense-making. Research on sense-making, as explained by Thomas and Cook [22], can provide a theoretical basis for understanding many of the analytical reasoning tasks that an analyst will require to perform. The sense-making loop, as shown in Figure 7, follows the processes of:

- **Information gathering:** The individual gathers the first pieces of information related to the task at hand.
- **Re-representation:** The information is re-represented in a form that will aid the analysis.
- **Development of Insights:** Creation of knowledge insights by manipulating the representation.
- **Creation of knowledge result-product:** The individual reaches a knowledge product based on the knowledge insights gathered.

These processes can be either repeated or executed out of order. Instances of such processes can be found in every aspect of modern-day life as well as in commerce, education and many other sectors. As an example, let us consider the purchase of a new smartphone. The potential buyer will first gather information from online shops and retailers. The information collected can be re-represented by creating a table with columns for each of the attributes that are of importance. This table could also be manipulated by adding or deleting columns for features of the smartphone that may arise as being important or unimportant. The potential buyer can then contemplate this table, delete or highlight rows and eventually formulate a rationalized purchase decision.

Data Representations, Transformations and Management

Visual analytics is powered by the data that can emerge from a variety of sources. These data may be of various types, as well as dynamic, ambiguous, incomplete, or even conflicting. Many are the times that Visual analytics systems have to analyze data that their representations are

not in the appropriate format to be analyzed. In such cases transformations to appropriate forms are required in order for an analysis to take place.

Each type of data has multiple perspectives, each of which influences the data's representation and the subsequent visualization techniques used to visualize the data. Some of the most important aspects [22] that someone should take into account when dealing with data are the following:

- **Data type:** The type of the data at hand. The type can be numeric, textual, audio, images etc. The majority of those types require some form of transformation prior to their analysis. These transformations may vary per type and even per different variations of the type. In the case of textual data, for example, they could range from plain text to categorical data and even dates. If analysis on plain text should take place, then the semantic meaning of the data should be evaluated and appropriate transformations to be formulated so that a meaningful visual analysis could be created.
- **Level of structure in the Data:** Data can have great variance in structure. Data can be structured, semi-structured or completely unstructured. The term "unstructured" refers to data that lack clear patterns that may be utilized for automated selection of the appropriate representation (for example, plain text data). In general, data structure may reveal how data are organized and serve as a foundation for building efficient data representations for later visualizations.
- **Geospatial or Temporal characteristics:** Data can also have references to space and time. Geospatial data are data associated with geographical locations or regions. This type of data could range from single geographical coordinates to ZIP codes, street names, municipalities, countries or some generic geographical region indicated by a group of coordinates. Depending on the format, a system should create appropriate transformations to an acceptable format for analysis. Temporal data, on the other hand, are usually textual data containing dates, timestamps or some form of time period description. If a temporal analysis has to be applied to data containing such information, then the system has to be able to transform these textual data to date formats that could

be utilized by a temporal analysis visualization. In some cases, temporal data could be combined with corresponding geospatial data, thus creating a new type of data called Spatio-temporal data. The analysis of data with references to both space and time is a challenging research topic since visualization techniques for visualizing geospatial and temporal data need to be combined in a manner that could help an analyst reach meaningful insights.

To create a suitable representation of data with potential characteristics such as those listed above, appropriate transformations must occur both during data ingestion and during data analysis. Only after such transformations have occurred and appropriate data formats have been reached and evaluated can the appropriate visualization techniques be chosen.

Another important aspect of data management in Visual Analytics is the storage and querying of the data. Simple ways of storing data like excel sheets and text files are not adequate enough for the necessities of modern-day Visual analytics systems. Over the past decades, relational database management systems (RDBMS) have dominated the market. The reason for their massive success was that with the use of the widely accepted SQL query language they allowed the accessing of large volumes of data in a controlled and managed manner.

Over the last two decades, and with the advent of web 2.0, the need for large scale databases due to large amounts of data, data structure, and application scale led to widespread industry adoption of NoSQL database solutions [22]. NoSQL (standing for "Not Only SQL") refers to a broad and increasingly known range of non-relational data management systems in which databases are not built primarily on tables and do not use SQL for data manipulation [23]. Some of the advantages of such databases are that they are distributed and capable of facilitating large scale data storage over a large number of servers which as a result increases the traffic that can be handled. They can also support parallel processing over multiple servers, while at the same time offering strong consistency meaning that all clients see the same version of the data. Such databases can also guarantee high availability with near-zero downtime and even partition tolerance. However, due to the consequences of the CAP theorem [24], high availability and partition tolerance cannot be provided concurrently, so a choice between the two is required

depending on the implementation. Finally, they are also able to deal with data of various structures and forms much more easily than relational databases thus enabling the support of the development of new application paradigms.

Data management is vital in Visual Analytics. Data should be examined and visualized based on their properties in order for the analyst to produce meaningful results. Furthermore, no application or system can succeed unless data is queried and supplied in a consistent, effective, and efficient manner, and new technologies in data management and databases can greatly assist in this regard.

Data visualization, issues and approaches

The ability to effectively present the outcomes of a data analysis is critical for any analyst or reviewer seeking to get insights from the data at hand. The field of data visualization is concerned in developing efficient methods of mapping between the actual data and graphic elements that are easier for humans to perceive and understand.

In the literature there are various methods for visualizing data [25, 26, 27]. These visualizations are usually either a graph or a table that can be classified in three groups, as static, dynamic or interactive.

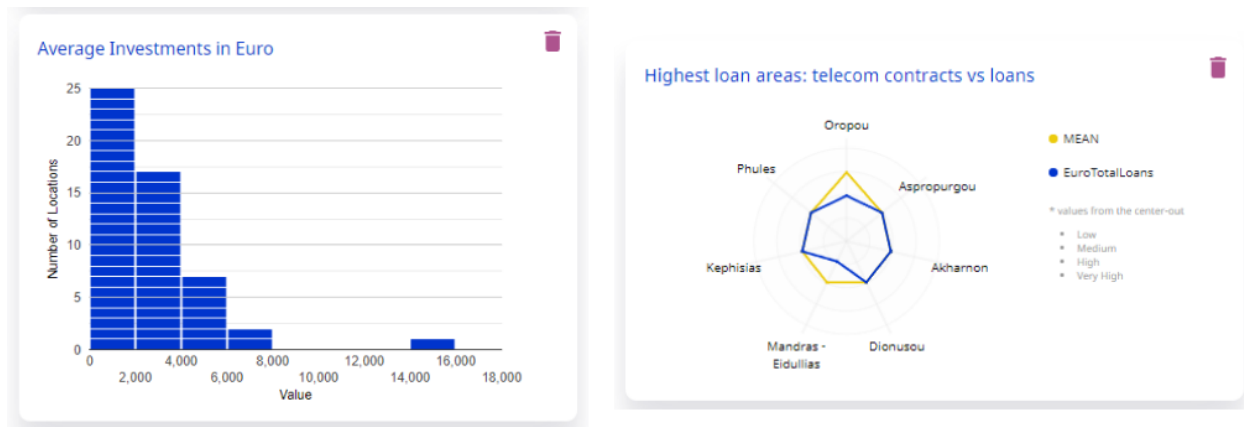


Figure 8 Examples of visualization techniques, a histogram diagram on the left and a radar chart on the right

Some of the more widely used visualization techniques are the following:

- **Line chart:** The type of graph that can represent the relation of one variable (x-axis) with another (y-axis) by depicting information as a series of points connected by straight line segments. This visualization can help with the tracking of changes that happen to the values of either variables and compare them at the same time.
- **Bar chart:** Bar charts can represent comparisons among discrete variables or values. Specifically, each variable or value is displayed as a bar (vertical or horizontal) in the chart and the height (or length if horizontal) given to the bar represents the value of the variable. In this way, substantial variance between variables or values can be identified.
- **Histogram:** This type of graph can indicate the variance in value density of an attribute. This is achieved by dividing the entire range of values of the attribute to smaller ranges, called bins. These bins are usually depicted as adjacent, consecutive, non-overlapping and of equal size bars. The difference between a histogram and a bar chart is that a histogram can only represent a single attribute on the x-axis and the data should be continuous. On the other hand, a bar chart can represent more than one attribute and the data representation can be non-continuous if needed (see Figure 8).
- **Scatter plot:** This type of graph uses the cartesian coordinate system to represent the values of variables. Each point on this graph has x and y coordinates that determine its position on it. For more than two variables to be represented, different coloring on the points can be used. A scatter plot is used to highlight the direction and linearity of the dependence between variables. A scatter plot can also be used for the representation of multidimensional data by adding another dimension. This 3D expansion of the scatter plot is called regressive cube.
- **Pie chart:** Pie charts, also called circle graphs, are a circle that is divided to sectors with each one representing a proportion of the total values of an attribute. Each sector can represent a different characteristic of an attribute and is usually color coded. Similar sectors, semantically, usually have similar hues of a color and different ones have different colors. Each sector also usually has a percentage value displayed over it describing its proportion compared to the total quantity. From a pie chart an analyst can perceive the proportions of different characteristics of an attribute.

- **Tree map:** This type of map is mainly used for displaying hierarchical data. It uses nested rectangles that reside within the main rectangle in order to represent the branches of a hierarchical tree. These rectangles vary in size and color depending on the size of the parameters they represent. This visualization can accurately illustrate the relation between hierarchical data, usually over a specific time period.
- **Radar chart:** Radar charts, or spider graphs, can represent multivariate data on at least three variables that are represented by axes that start from the center of the diagram (see Figure 8). The length of each axis is proportional to the magnitude of the variable for the data point in relation to the maximum magnitude of the variable across all data points. A line is created linking the data values for each axis thus giving the plot a radar-like appearance. This chart is usually used for detecting outliers or clusters of similar observations.
- **Box plot:** This type of plot is mainly used for visualizing the distribution of numerical data and skewness. It achieves that by displaying the minimum, first (lower) quartile, median, third (upper) quartile, and maximum values of an attribute as well as outliers. The first quartile, median, and third quartile usually form a box in the graph, which is the reason the plot was named “box”. The minimum and maximum values are depicted as lines that extend from the box and up to their value. The outliers of the plot are usually displayed as points on the graph.

The visualization techniques discussed above are just some of the available techniques with others being: violin charts, donut charts, area charts, bubble charts, flow charts, Gantt charts, Sankey diagrams and many more, all related to a specific type of data and revealing different kinds of insights through their visualizations.

When choosing the right visualization techniques to be applied on the data, an analyst or a developer of a visual analytics application should always take into consideration some of the most common drawbacks related to some of the techniques mentioned previously. Those drawbacks almost always arise from the fact that big data visualization and big data in general have some elements of uncertainty like volume, velocity, variety, veracity and value [28]. In terms

of visualizing big data, studies have shown that some of the most common problems are the following [26, 27]:

- **Visual noise:** This problem usually emerges when visualization approaches are selected that attempt to present all of the data under study at once. As a result, the screen becomes cluttered, with the user attempting to comprehend this representation only being able to perceive one large spot rather than the points that comprise that spot. For the analysts, this means that they cannot get any useful information or insights.
- **Large image perception:** Some naïve methods to dealing with visual noise were to simply raise the resolution or size of the screen itself. This approach, however, yields new issues. In general, humans have a limited ability to perceive large data visualizations. As a result, shifting the problem of visual noise to a larger screen tends to produce perception issues since, while the information can be presented, a human analyst will struggle to interpret it.
- **Information loss:** One method for dealing with both visual noise and large image perception is to limit the amount of data shown. Although this may successfully address the aforementioned issues, it may also result in information loss. In particular, if data are aggregated and filtered in order to exclude records from the dataset that are thought to be highly related to one another, this may result in an analyst not noticing some valuable information that could be hidden in them.
- **High performance requirements:** The need to visualize large volumes of data usually comes at a performance cost. This cost may vary depending on whether the visualizations are static or dynamic. Modern day computer systems as well as advancements in storage options have helped address this issue, still some filtering or aggregation is needed for visualizations that require large volumes of data.
- **High rate of image change:** This issue can occur when data is constantly changing in front of the screen of analysts without them being able to understand or react to them. Such visualizations cannot easily be fixed because just a simple decrease on the changing rate of visual images won't be enough since the limiting factor is mainly the reaction speed of the individual.

Visual Analytics greatly depends upon the successful representation of data. In the literature there are many visualization techniques that aim at visualizing various types of data. When an analyst or a developer has to make decisions about which visualizations are best suited for the task at hand, they should take into consideration many factors. These factors range from the type of the data to the limitations that their great volume might cause to both the analyst trying to understand them as well as the system trying to generate them.

2.2.2 Visual Analytics Process and Knowledge Generation

The Visual Analytics process has been described by Keim, Daniel, et al. [18] as the process that combines Visual Analytics methods through human interaction in order to gain knowledge and insights from the data at hand. This process is in the form of a closed loop where the main nodes are the data, the visualizations, the models and the knowledge (see Figure 9).

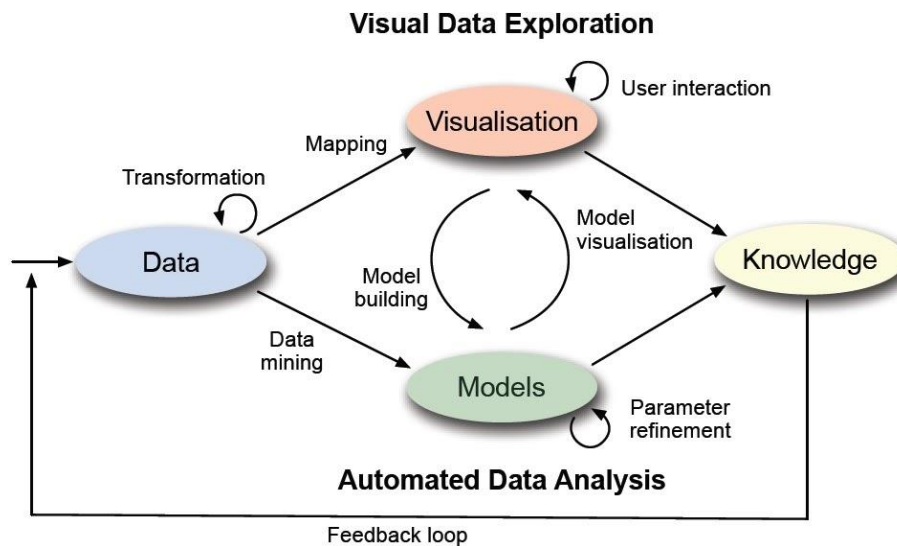


Figure 9 The Visual Analytics Process [18]

Data can be considered as the starting point of any Visual analytics process. These describe information that can have several peculiarities as mentioned in previous sections of this chapter and the goal for every analysis is to reach from raw data to new knowledge and insights. In this stage of the process, the data, that might be heterogeneous or of different types, need to be transformed into appropriate forms that can be utilized from both automatic and visual analytics methods.

When the data are in the proper format, the analyst can choose to either conduct an automated data analysis or use visual data exploration in order to get to the desired knowledge. With an automated data analysis, the analyst can use data mining methods and the KDD process [29] in order to generate models. These models can range from calculating a single number to detecting clusters of interesting data. In this stage, visualizations can help the analysts with the fine tuning of the parameters of the models and judge whether or not the chosen analysis algorithms can serve their purpose or need to be changed and the model to be altered. The analyst can also use visual data exploration to generate visualizations, search through data and detect relationships within them, which can then aid in the formulation of automated analysis models. The desired results and new knowledge can be obtained (as seen in Figure 9) from either the visual exploration or the automated data analysis route. Then, based on the analysts' iterative interaction with the system, this new knowledge can be used as input data to the loop, thus refining the insights derived from the data.

Based on the aforementioned framework, Sacha et al. [30] introduced the knowledge generation model for visual analytics, which added individual steps to further describe the overall visual analytic process.

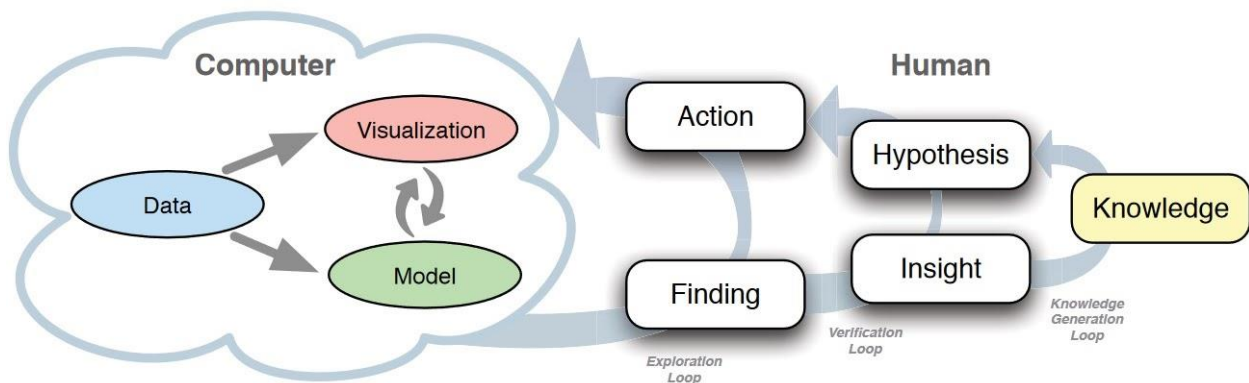


Figure 10 Visual Analytics process and the Knowledge generation model [30]

This approach consists of computer and human parts since the visual analytics process strives to employ both the human's ingenuity for the detection of subtle or hidden relationships between data, and the computers' capacity to deal with large amounts of data.

The computer part, as before, consists of the models, which deal with the analysis of the data through KDD approaches and data visualization based on the specified data models as well as visual data exploration. The human part consists of two loops, the exploration and verification loops, which aim to generate new information and insights, which will then be utilized to iteratively improve the quality of the final knowledge produced from the entire process.

The exploration loop mainly focuses on the way in which analysts are interacting with a visual analytics system and specifically how they are creating new models and visualizations through the data. Specifically, the loop is composed of the Action and Finding processes. In the Action process, a taxonomy is given that describes the interaction between the analyst and the system:

- **Preparation:** Actions that are dealing with the data selection and the required transformations in order for them to fit the scope of the analysis.
- **Model building:** Here, KDD and data mining is used for the creation of models.
- **Model usage:** Application of the models created from the previous step to the data.
- **Model-visualization mapping:** Mapping of the created models to visualizations.
- **Manipulation:** Manipulating the created visualizations in order to spot and highlight interesting data.

The Finding process, after the interactions mentioned above have been performed, is concerned with the observation of the results by the analyst. At this point, any meaningful observation and feedback from the system can help the analyst adjust and improve both the models as well as the visualizations and thus enhance or discover insights.

The verification loop, which has as an objective to lead the exploration loop, can help with the validation on hypotheses or help create new ones. It consists of two processes, the Insight and Hypothesis. Hypotheses are the assumptions relative to the problem that is being analyzed and the aim of the visual analytics process is to either prove or disprove them. Therefore, the

hypotheses can be considered as the starting point from which the analysts can then use the exploration loop in order to investigate them. If the exploration loop leads to insights that can support the hypothesis, then this hypothesis can be considered as valid and as a new addition to the knowledge derived from the system. If the exploration loop reaches insights that disprove the hypothesis, then the hypothesis can be adjusted or removed completely and a new one can take its place at the loop. In this way, the Visual analytics process becomes an iterative process for knowledge generation.

2.2.3 Data Visualization Applications and Limitations

Data visualization tools and applications come in a variety of forms in the literature. From libraries in programming languages like R, Python and JavaScript to applications for non-computer scientists like Excel. Our research will primarily focus on applications that can provide their users with the ability to analyze and explore data without the need for programming experience, while also providing powerful calculations and statistical summaries.

Some of the characteristics that are expected from such data visualization applications are the ability to select attributes and specific statistics for visualization, the ability to create dashboards that can facilitate those visualizations, and compatibility with storage options. Dashboards, in particular, are where most of the analysis takes place. They are composed of data visualizations that can provide an overview of the data as well as filtering techniques for limiting the data under analysis. Analysts can create and customize visualizations as well as filters, allowing them to explore the data by focusing on specific aspects of the dataset's attributes. Aside from customizing dashboard visualizations and filters, another method of allowing data exploration is to apply drill down or roll up processes to the data.

By interacting with the visualizations and filters provided by the dashboard, an analyst can drill down from a broad grouping of data to a more detailed and granular grouping of data within the same dataset, thus stepping down from one level of a data hierarchy to the next. Data roll up can be considered as the reverse process of drill down, where an analyst can step up from a specific data hierarchy level to a more general one thus investigating data in progressively less detailed levels. For example, in the case of geospatial data, one hierarchy could be that of country, county,

CHAPTER 2. LITERATURE REVIEW

municipality. In this case, a drill down process would be shifting from data involving counties to data involving municipalities within a county, and vice versa for the roll up process.

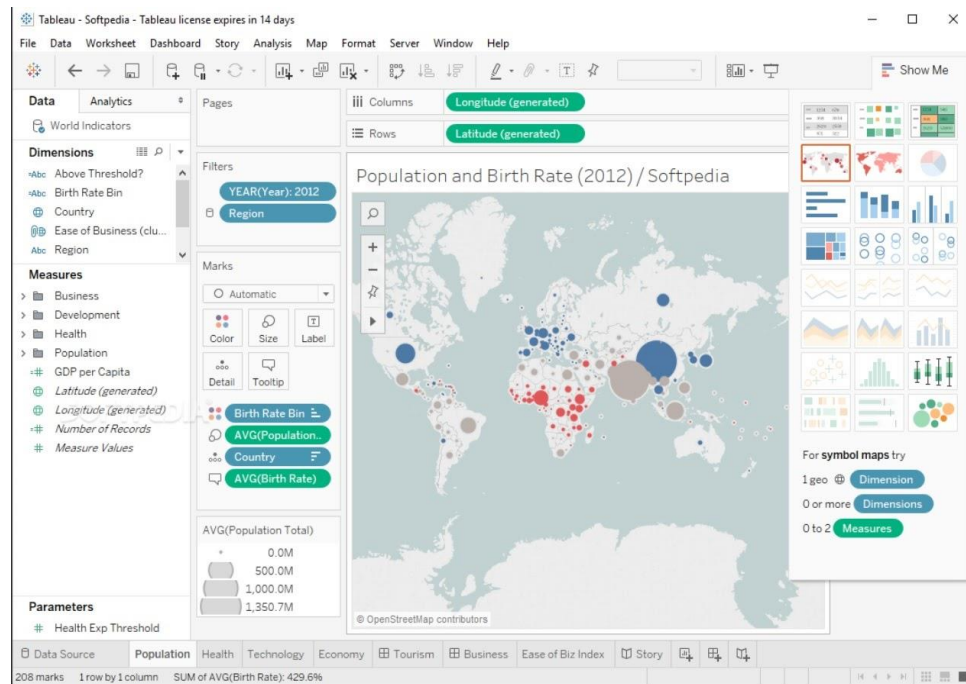


Figure 11 Tableau Desktop

There are numerous applications and tools available on the market and in the literature that can provide the majority of the above-mentioned functionalities. Tableau [31], though, is one of the most popular applications of its type, considered as the state of the art in the specific domain.

Tableau is a data visualization software that can be either deployed on the cloud or on-premises and offers a variety of data visualization capabilities that an analyst can access through dashboards in order to explore data and discover insights. Tableau can fetch data from various resources like MongoDB and Elasticsearch [32]. After the data is loaded, users can select attributes, and the system will automatically select a default data representation based on the attributes selected. Users can enrich the visualizations by adding new attributes to them. There is also a section with a gallery of several visualization templates that users can use, with those that are not available due to the current attribute selection automatically grayed out.

Another well-known software is Power BI [33]. Power BI can support the creation of interactive dashboards and reports based on data from various data sources. After the users have inserted

the data, they can transform them and then choose the attributes they want to visualize. Like Tableau, there is a gallery of visualizations the users can choose from, as well as filters they can apply. Unlike Tableau though, Power BI is limited to the volume of data it can import (cannot accept file sizes above 1 GB) which might pose an issue for users with datasets above this size.

One area where both of these software solutions provide options but are generally not intuitive is in data drill down and roll up processes. They specifically provide this functionality through the use of various filters and selections that require users to go through the values of various attributes and select which ones they are interested in investigating as well as having to manually setup the granularity of data drilling depending on their action. Furthermore, this functionality is typically limited to a single map or chart, and data drilling actions do not propagate to the rest of the dashboard's visualizations without extensive customization, potentially limiting the insights that a user can discover.

Chapter 3

Methodology

The proposed methodology aims at addressing the open issues analyzed in the previous chapters and specifically in sections 2.1.8 and 2.2.3. In detail, it provides a unified framework aimed at the analysis, anonymization, visualization, and exploration of Big Data while ensuring security and privacy. In this chapter, the proposed framework is detailed, including the pertinent applications that have been implemented on top of this framework, elaborating on the functionalities that they provide.

3.1 Framework

The proposed framework addresses two main issues concerning the data exchange cycle between data owners and digital data marketplaces (as seen in Figure 12). These are the issues of protecting the privacy of individuals included in datasets about to be traded, as well as the need to gain new insights through the correlation and analysis of already owned data with data obtained from such marketplaces.

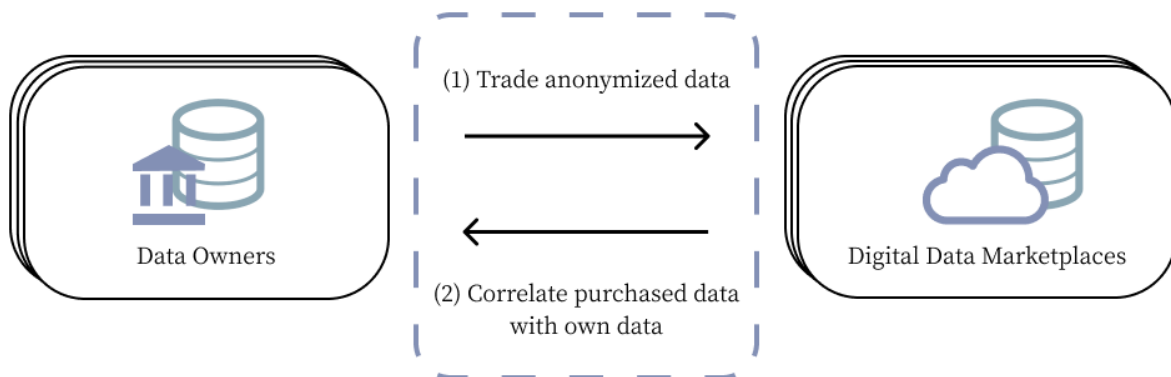


Figure 12 Proposed Framework

In particular, to protect individuals' privacy, under this framework, data owners must ensure that the data about to be traded in digital marketplaces is of low risk of being de-anonymized and, if not, anonymizing it prior to trading. To that end, data owners should conduct de-anonymization risk analyses in order to analyze the risk of identification of individuals from their sets of data and, afterwards, anonymize the datasets that were identified as of high risk from these analyses. In this way, data owners can trade data in a privacy-preserving manner while ensuring their compliance with data regulations such as GDPR and also contributing to increased trust in such marketplaces.

Concerning the second issue raised by this framework, data owners who engage in trading activities with digital data marketplaces typically intend to correlate data from various business sectors with their own in order to improve several aspects of their businesses. To that end, data owners should be able to conduct targeted analyses and explore these datasets through the use of data visualization and visual analytics techniques. Such techniques should include the use of dashboards that can adapt to the user's analysis framework requirements as well as data drill down and roll up functionalities that could be applied on combinations of datasets. In this way users will be able to explore data and gain new insights that they would not have gotten otherwise.

By addressing the aforementioned issues, as described above, the proposed framework offers a unified approach to dealing with both the privacy risks of datasets about to be exchanged and the analysis of datasets obtained from digital data marketplaces. Data trading can thus be done in a trustworthy and privacy-preserving manner, while also allowing data owners to gain high-value insights from the analysis of such data.

3.1.1 Framework Realization and Specifications

The presented framework employs two applications that provide the necessary intuitive User Interfaces (UIs) so that the users are capable of addressing the above mentioned needs. The first one is a toolkit that provides a complete solution for facilitating de-anonymization risk analysis as well as anonymization functionalities to assist data owners in analyzing the risk of identifying

individuals from their sets of data, as well as providing them with the ability to anonymize the respective datasets.

The second application provides Big Data visualization functionality allowing users to create targeted analyses on datasets by providing the ability to combine datasets and conduct concurrent analysis of both datasets through interactive smart dashboards. These dashboards are able to adapt to the user's analysis framework requirements and provide data drill down and roll up functionalities based on the type of data being analyzed.

The specifications for both applications that realize this framework were elicited through questionnaires as well as numerous focus group meetings with data analysis, marketing, and technical experts from telecommunications and financial industries in the context of the Horizon 2020 EU project TRUSTS [34]. These specifications were later compiled into task analyses, which served as the foundation for the design and development of both applications.

3.2 Deanonimization Risk Analysis Application – DaRAV

3.2.1 Application Description

DaRAV (De-anonymization Risk Analysis through Visualizations) is a toolkit designed and developed to provide a platform for conducting de-anonymization risk analyses in given datasets of various data types, as well as creating anonymized versions of these datasets based on the results of the analyses. It offers a complete solution for accommodating either risk analysis or anonymization modules, with support for dataset importing, risk analysis or anonymization process configuration, queueing and parallel execution of such processes, as well as storing and visualizing risk analysis results.

Supported Datasets

In the context of the present application, the term “dataset” is used for the representation of a file in the user’s machine. Each dataset has a set of metadata information that go along with it. Following are the data that need to be filled out by the user while importing the dataset in to the application:

- **File location and separator/delimiter**

The user can specify the file that they want to import from their local filesystem. The file has to be in .CSV format. They also have to set the type of delimiter (comma, semicolon or tab) used in the particular .CSV file.

- **Data type**

Data type is referring to the type of the data, e.g., tabular, aggregated, location data, etc. contained in the .CSV file specified.

- **Title and short Description**

The user can also set a title for the dataset as well as a short description for further explanation of the data contained in it.

The data and metadata related to the datasets imported to the application are all stored locally, in the user’s machine which is running the application.

Risk Analysis Modules

The application utilizes several de-anonymization risk analysis modules in order to compute and later visualize the risks that a dataset might have at leaking personal data. Specifically, six de-anonymization risk analysis modules are currently used. These modules consist of algorithms which were designed to address various types of de-anonymization threats based on the data type or its peculiarities. The data types supported are:

- **Tabular data:** data that are structured into rows (records), each of which contains information about an entity
- **Spatiotemporal data:** data with records that have references to both space and time
- **Textual data:** data that have records comprised of textual attributes e.g.: user comments or reviews
- **Data of Financial nature:** data that contain information about financial transactions
- **Aggregation-based data:** data that contain aggregate values, e.g.: sum, count, average, of an entity

For each of these types, a module can compute a risk analysis based on the parameters that the user has specified. Each module requires a different set of parameters to be entered into a form provided by the application when a user wishes to conduct a risk analysis process. The following are the risk analysis modules provided by the application, the modules/algorithms used for each one, the data they generate, and how they are visualized. The risk algorithms of these modules were implemented in the context of the Horizon 2020 EU projects TRUSTS [34] and Safe-DEED [35].

K-Anonymity

This module can compute a tabular dataset's compliance to k-anonymity and thus quantify its risk for de-anonymization. On the risk analysis page of the application, users are given a form in which they must select which of the dataset's attributes (columns) should be considered as Quasi Identifiers, as well as the k parameter of the k-anonymity for which they wish to compute the analysis. Then the k-anonymity module computes the percentage of compliance for each of the unique combinations of the dataset's QIs.

De-anonymization Risk Analysis

Tabular Dataset

Tabular Dataset (Adult Dataset)

Risk Analysis Method

k-anonymity

Results Chart

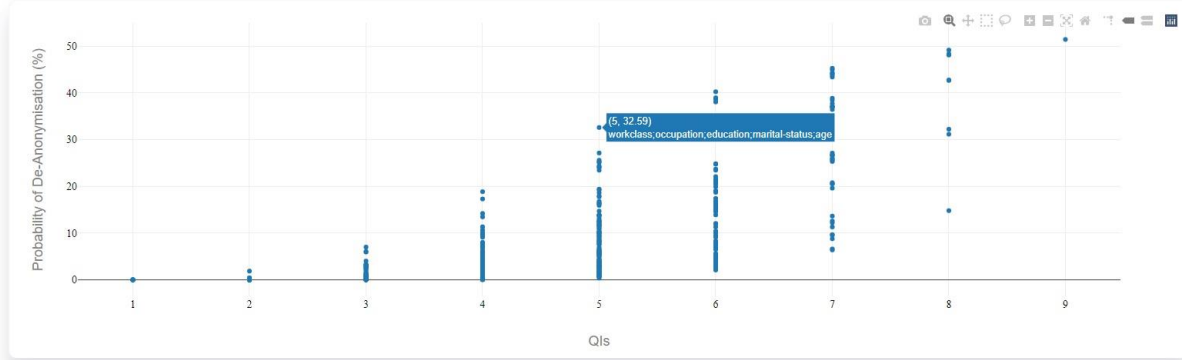


Figure 13 Results Page – The results of a risk analysis process with k-anonymity ($k=2$)

In Figure 13, the results of such a process are being visualized in the form of a plot. The x-axis refers to the number of QIs in a combination and the y-axis refers to the probability of de-anonymization, given that an adversary possesses the values of $k-1$ individuals of any of the combinations of QIs. For example, as seen in the figure, if an adversary possesses information about the “workingclass”, “occupation”, “education”, “marital-status” and “age” attributes of an individual ($2-1=1$), then the adversary has an 32.59% probability of de-anonymizing an individual.

L-Diversity

This module can compute a tabular dataset’s compliance to l -diversity and thus quantify the risk of an adversary finding out the value of a sensitive attribute. On the risk analysis page of the application, users are given a form in which they must select which of the dataset’s attributes (columns) should be considered as quasi identifiers and, in addition to the form provided for k-anonymity, they also have to select the sensitive attributes of the dataset. Lastly, they also have to specify the l parameter of the l -diversity for which they wish to compute the analysis. Then the l -diversity module computes the percentage of compliance for each of the unique combinations of the dataset’s QIs and sensitive attributes.

CHAPTER 3. METHODOLOGY

De-anonymization Risk Analysis

Tabular Dataset

Tabular Dataset (Adult Dataset)

Risk Analysis Method

l-diversity

Results Chart

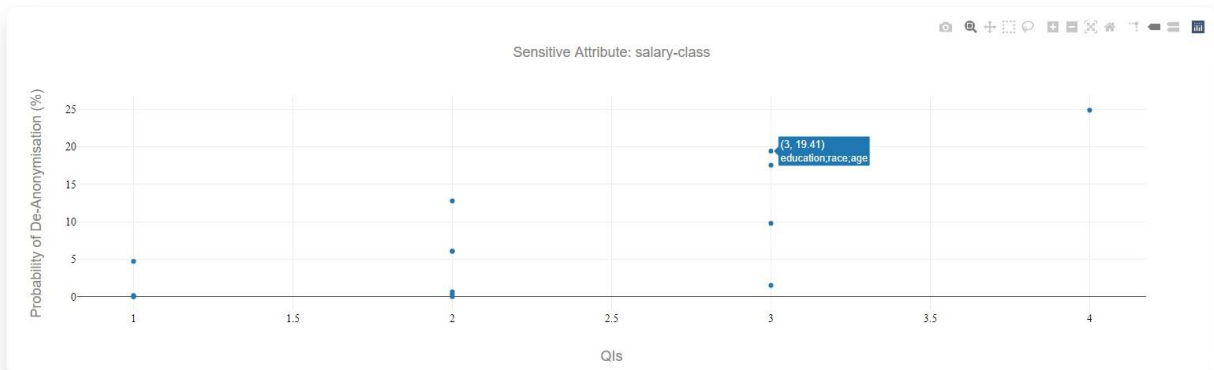


Figure 14 Results Page – The results of a risk analysis process with l-diversity ($l=2$)

In Figure 14, the results of such a process are being visualized in the form of a plot. The plot is similar to that of k-anonymity with the difference that the probability of de-anonymization now considers the probability of an adversary finding out the value of the specified sensitive attribute, in this case “salary-class”.

Location

De-anonymization Risk Analysis

Spatiotemporal Dataset

Spatiotemporal Dataset (Gowalla Dataset)

Risk Analysis Method

Location

Results Map

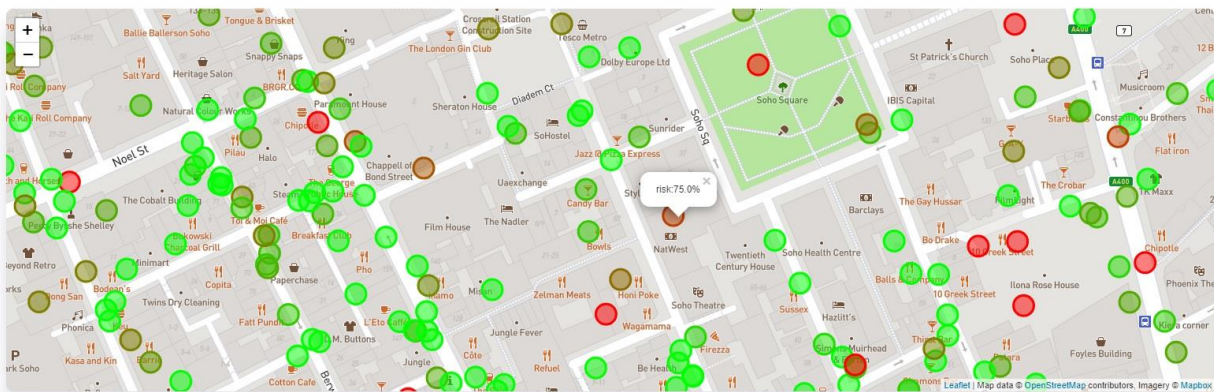


Figure 15 Results Page – The results of a risk analysis process of location data for $w=1$, $r=200m$, $t=6$ hours

Spatiotemporal data are data that contain information about an entity's location and time. This module computes the probability of singling out an individual by calculating the compliance of spatiotemporal records to the following privacy notion:

w other individuals within a radius r , within a timeframe t

In detail, if an individual is located at location x at time y , and there are at least w other individuals within a radius r , within a timeframe $y \pm t$, then the location x at time y is considered safe.

In Figure 15, the results of such a process are being visualized in the form of a map. Each point on the map represents an individual's location. The color scale of the points is from red to green. The greener a point is, the safer the respective location is. On clicking a point on the map, a pop-up is displayed with the point's probability of de-anonymization. In this way, the data owner can have an overview of the privacy-utility trade-off for an anonymization method that makes the data conform to this privacy principle.

Textual

De-anonymization Risk Analysis

Textual Dataset

Textual Dataset (Amazon Reviews)

Risk Analysis Method

Textual

Results Chart

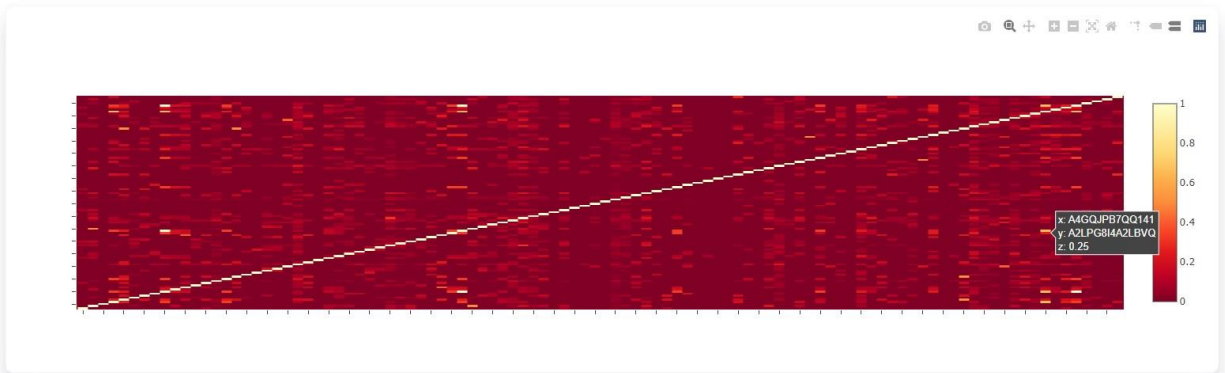


Figure 16 Results Page – The results of a risk analysis process of textual data

Textual data are records that contain information that individuals have provided in written form. Examples of such data are datasets related to user reviews of products or services, as well as user search logs on web sites. This module calculates the de-anonymization risk of a dataset by

measuring the uniqueness of words used by individuals. The results of this module are based on the similarity of texts of two individuals which is computed with the use of the Jaccard Similarity:

$$J(u_i, u_j) = \frac{words_{u_i} \cap words_{u_j}}{words_{u_i} \cup words_{u_j}}$$

The Jaccard Similarity calculates a score between 0 and 1 for every pair of individuals from the dataset. The results are visualized in a heatmap plot as seen in Figure 16. The x and y axes correspond to individuals. Each tile of this heatmap contains the identifiers of two individuals as well as the results of the Jaccard Similarity between the texts of the two. The color scale of the tiles is from red to yellow. The yellower the heatmap gets, the higher the de-anonymization risk.

Invoices

Financial transactions data, like invoices, are data that contain information about payments that individuals made or received at a specific time in the past.

De-anonymization Risk Analysis

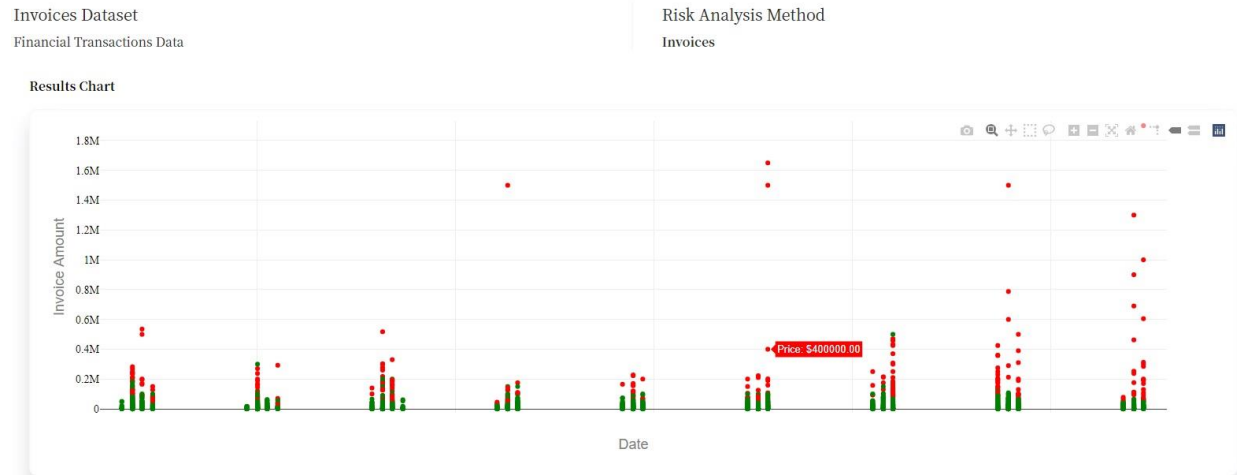


Figure 17 Results Page – The results of a risk analysis process of transactions data ($w=1$, $a=1000$, $t=1$ month)

This module computes the probability of singling out an individual by calculating the compliance of transactions records to the following privacy notion:

w other individuals having a transaction amount within *a*, within a timeframe *t*

In detail, if an individual has paid an amount x at time y , and there are at least w other individuals having a transaction amount $x \pm a$, within a timeframe $y \pm t$, then amount x at time y is considered safe.

In Figure 17, the results of such a process are being visualized in the form of a point plot. All the unique data points of amount (x-axis) and date (y-axis) are colored green if they comply with the specified privacy notion and red otherwise. The output of this risk analysis module reveals the outliers of the data, i.e., the individuals with distinct transactions. Additionally, the output can help the data controller decide the generalization hierarchies (i.e., the binning of the amounts) in the case of anonymization, or the aggregation levels in case of aggregation.

Aggregated

Aggregated data are data that contain summed up values about individuals, e.g.: sum, count, avg. Such data usually don't pose a de-anonymization threat, but if there is a sensitive aggregate attribute within the data and the aggregate values are low, then a privacy breach may occur.

De-anonymization Risk Analysis

Aggregated Dataset

Aggregation based data

Risk Analysis Method

Aggregated

Results Chart

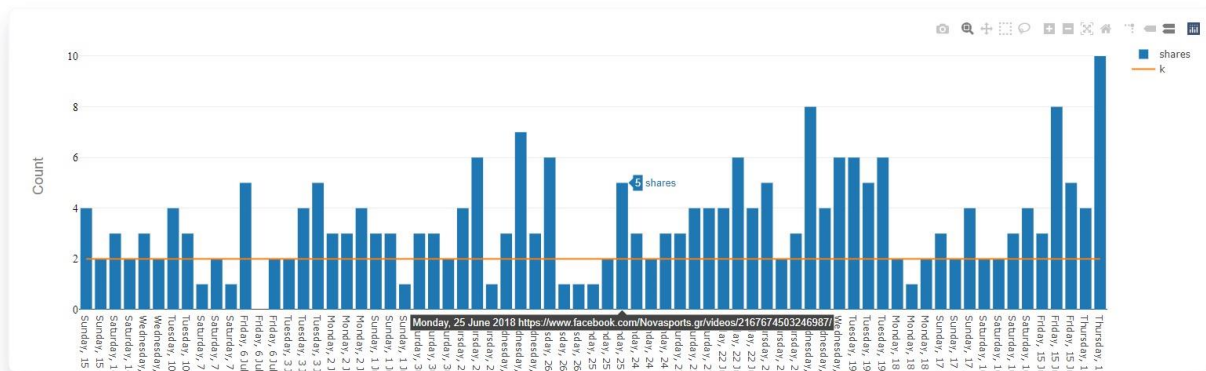


Figure 18 Results Page – The results of a risk analysis process on aggregated data

In Figure 18, the results of such a process for the attribute “shares” of a social media related dataset are visualized in the form of a bar plot. Each bar represents the total records of the dataset containing the same value on their “shares” attribute. The plot also has a horizontal line (k) representing the minimum acceptable value (total records) of the aggregation attribute. In

this case, if another sensitive aggregated attribute exists, the records that include the attribute "shares" and are below the line should be removed.

Anonymization Modules

In addition to the de-anonymization risk analysis methods, the application also provides anonymization methods that the users can utilize. Users can use the information provided by the risk analysis modules results in order to choose the appropriate anonymization method as well as the appropriate parameters and attributes of a dataset. Currently, the application supports two methods: k -anonymity and l -diversity.

To perform k -anonymity anonymization on a dataset, users are presented with a form in which they must select which attributes (columns) of the dataset should be considered as Quasi Identifiers, as well as the k parameter of the k -anonymity. They must also choose a .json file from their filesystem that defines the generalization hierarchies of the dataset's attributes. A similar input is required for l -diversity, where the QIs and sensitive attributes of the dataset, as well as the l parameter and a .json file containing the generalization hierarchies, must be specified.

After the completion of an anonymization process, the users can save the newly anonymized dataset on their filesystem.

3.2.2 Implementation

Application Architecture

The application was developed in a loosely coupled manner, where every component is being a standalone entity in a separate docker image. All the components/images needed for the application are deployed through a single docker-compose file.

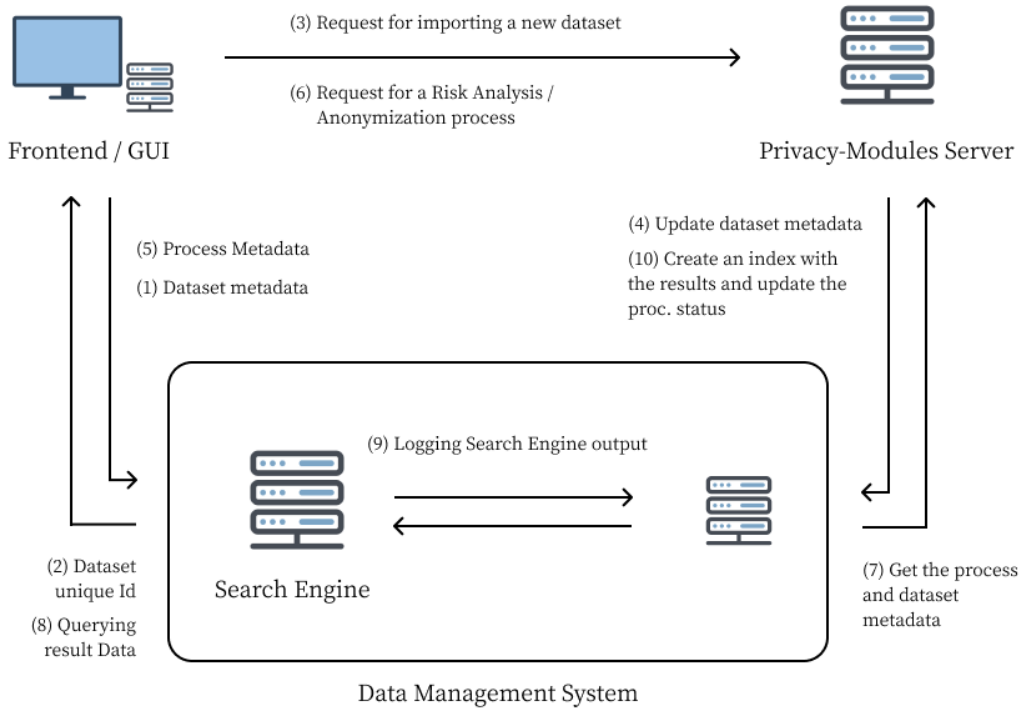


Figure 19 DaRAV Application Architecture diagram

The main components of the app's architecture are the following:

- The application frontend/GUI
- The Privacy-Modules Server, which undertakes the importing of the dataset to a local SQLite database and execution of the risk analysis and anonymization algorithms as well as the importing of the results in the Search Engine
- The Data Management System which stores all the information transacted between the components (metadata), the resulting data of each process as well as the logging of the app's components outputs

In Figure 19, the diagram of the architecture depicts the flow of information between the API's of the individual components. These workflows are mainly triggered by the user's actions (as they will be described in the Using the Application section, 3.1.3). Following is a brief description of each one:

1. Dataset Metadata

The user chooses a dataset to be “uploaded” (imported) and enters relative metadata: CSV delimiter character, title, sort description and type of the data, through the upload page's form.

2. Dataset unique Id

Through an AJAX Request the frontend/GUI sends these metadata to the Data Management System, where a new document (that contains them) is created in the index responsible for storing the metadata related to datasets.

3. Request for importing a new dataset

With the dataset's metadata already collected, an AJAX request is sent to the Privacy-Modules Server's API so that the import dataset process can start.

4. Update dataset metadata

The Privacy-Modules Server updates the dataset's metadata document with the import progress as well as other metadata. After the dataset import is completed, the user can initiate a risk analysis or an anonymization process for this dataset.

5. Process Metadata

From the app's GUI, a user can initiate a process (risk analysis or anonymization) for a specific dataset and fill out the necessary parameters in the respective form that is provided. After all the necessary metadata is set for the process, a new document is created in the index responsible for that type of process in the Data Management System.

6. Request for a Risk Analysis / Anonymization process

After the document containing the metadata of a process is created in the Data Management System, an AJAX Request is sent to the API of the Privacy-Modules Server containing the unique Id of the process that needs to be initiated.

7. Get the process and dataset metadata

With the use of the process' unique Id, the Privacy-Modules Server retrieves the metadata from the Data Management System and initiates the corresponding task based on the parameters specified.

8. Create an index with the results and update the process status

When the Privacy-Modules Server has the process' results ready, it creates a new index in the Data Management System and stores them. When this operation is done, the server updates the status of the process to "Completed" and the user is able to review the results from the process' page.

9. Logging search engine output

For the purposes of logging the output of the Data Management System (Elasticsearch node), Logstash will be used.

Technologies used

For the frontend / Graphical User Interface (GUI) of the application, the framework used is Angular v10 [36]. Angular is a widely used, TypeScript-based, open-source web app framework capable of creating robust web applications. For the geographical maps, charts and graphs offered by the application, various libraries such as leaflet [37] and plotly [38] are used. The Data Management System is based on the Elasticsearch [32] search engine. Elasticsearch comes as a part of the Elastic Stack which is composed, mainly, of three open source projects: Elasticsearch, Logstash, and Kibana. Elasticsearch is a distributed, free and open search and analytics engine for all types of data, built on Apache Lucene and is more than capable of meeting the storing as well as the communication requirements of the application. Logstash [37] is a free and open server-side data processing pipeline that ingests data from a multitude of sources, transforms it, and then sends it to the Elastic Stack. In the context of the application, Logstash is used to ingest to an Elasticsearch index the logs of the deployed Elasticsearch node.

The Privacy-Modules Server is created with Java and the Spring framework. Spring Boot [39] with the Maven Plugin [40] is used for the data import, the computation of the processes as well as the creation of the API. For storing and efficient querying of the data, an SQLite database [41] is used. The use of this database over the storing of the data to the Elasticsearch node was decided

since the only need for storing the datasets to the system was for the querying of the data for the purposes of conducting risk analysis or anonymization processes. Ingesting the data to the Elasticsearch node would mean that this data would have to be transferred through AJAX requests to the Privacy-Modules Server in order to be processed. The SQLite database allows the server to query data without the use of AJAX requests.

For the deployment of the application, Docker is used. Docker [42] is an open platform for developing, shipping, and running applications that enables the separation of the application from the infrastructure. As mentioned previously, all the components of the application are housed in separate docker images and deployed with a single docker-compose file. In this way, each component of the application is independent of the others yet capable of communicating with them through well-defined API's.

3.2.3 Using the Application

After the successful deployment of the application, a user can have access to the Graphical User Interface (GUI) by visiting a specified link on an internet browser. Upon entry, the user will be asked for their credentials in order to log in. After a successful authentication, the user is able to view all the imported datasets or choose to upload a new one. For each dataset, the user can initiate either a Risk Analysis or an Anonymization process. After filling out all the required parameters, the chosen process can commence. All the previously initiated processes, depending on their type, can be found in the respective processing queues. When a process is completed, the user is capable of viewing the results of the process in the respective results page. Following are the actions, as described above, that a user can accomplish through the GUI.

User Login and Application Layout

The user can log in to the application by entering their credentials. If the authentication process succeeds, they proceed to the Datasets Page. After a successful authentication, the user is directed to the Datasets page. All the pages offered by the application follow the same design pattern. This design pattern consists of three components: the horizontal navigation menu (on the left sidebar), the header and the content of the page (see Figure 20).

From the horizontal menu, the user can navigate through the main pages of the application, namely: Datasets, Risk Analysis, Anonymization and Processing Queue. At the top of each page, two buttons are located, the one for viewing the notifications of the user and the other for viewing the profile. The rest of the space, depending on the page, will display appropriate content.

Viewing the Available Datasets

In the Datasets page, a user can view all the datasets imported to the application. There are two viewing options: the card view (set by default) and the list view. The user can specify the desired view with the use of the buttons located at the top-left of the page (below the user profile button).

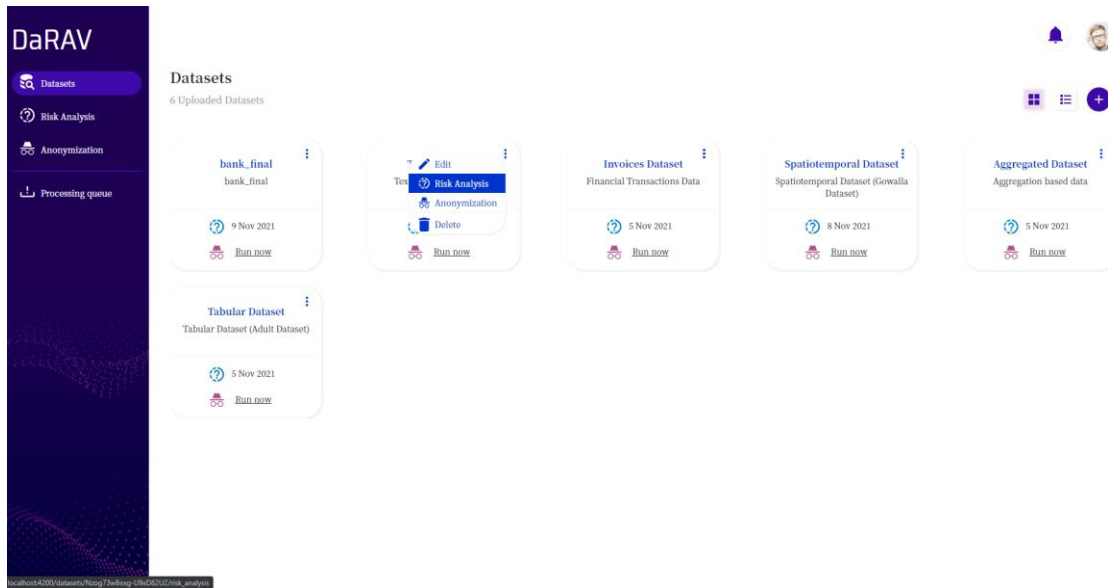


Figure 20 Datasets Page – card view

For each dataset, the title, description and last time of running a risk analysis or anonymization process are displayed. From the options menu (three blue dots) the user has the ability to edit, delete or initiate a Risk Analysis or Anonymization Process for a dataset. While a dataset is being uploaded, the user is only able to edit or delete it. After the uploading process is completed, the options for a Risk Analysis or an Anonymization are enabled.

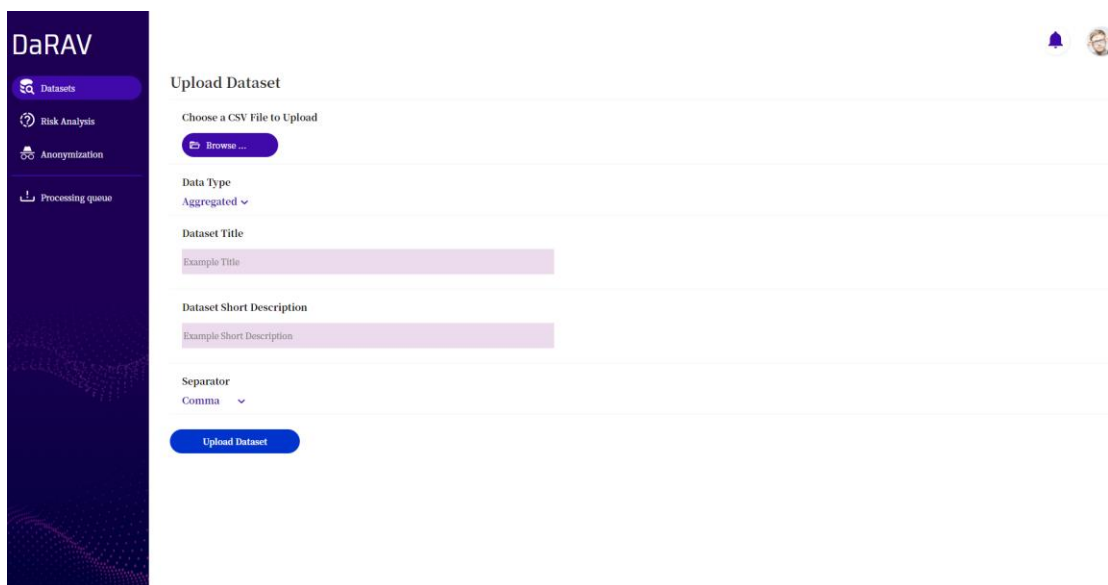


Figure 21 Upload Dataset Page

Uploading a new Dataset

A user can import a new dataset by selecting the “Upload Dataset” button located at the top-left of the Datasets page (below the user profile image). In the Upload Dataset Page (see Figure 21), a user can specify the .CSV file containing the dataset, the dataset’s title and short description, delimiter/separator of the .CSV file as well as the data type of the dataset. After filling out this information, the importing process begins. Until this process is finished, the user is only able to edit or delete the dataset. After the completion of the process, the Risk Analysis and Anonymization options are enabled.

Editing a dataset

After selecting the “Edit” option of a dataset, the user is directed to the dataset’s info page. In this page, all the basic information of the dataset is displayed. The user has the ability to edit the dataset’s title and short description by selecting the pen icon beside either the title or the short description.

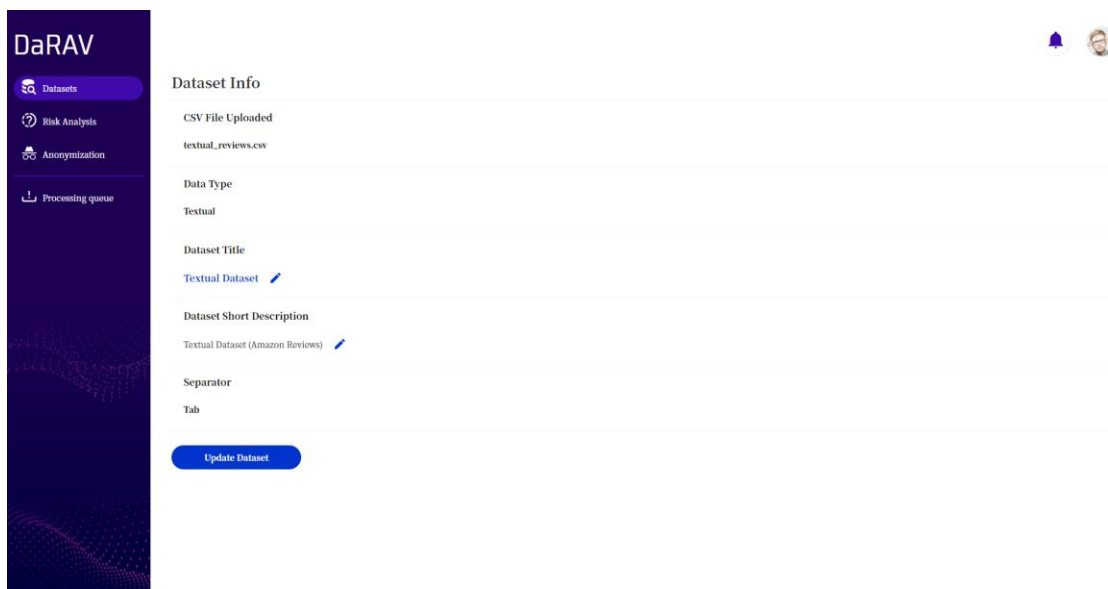


Figure 22 Dataset Info Page

Starting a new Risk Analysis or Anonymization process

To start a new risk analysis or anonymization process, a user has to choose a dataset from the Datasets page and then select the “Run Now” button, or, from the dataset’s options menu, the “Risk Analysis” or “Anonymization” option respectively.

DaRAV

Risk Analysis

Invoices Dataset
12 Columns | Financial Transactions Data

Risk Analysis Method
Invoices ▼

| Name ▼ | Individual Identifier | Invoice Date | Invoice Amount |
|------------------------|----------------------------------|----------------------------------|-----------------------|
| PI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Customer Code | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Order Number | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Order Entry Date | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> |
| Customer Wish Date | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Confirmed Delivery ... | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Requested Order Lea... | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Confirmed Order Lea... | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Processing parameters
Date format: yy.MM

Risk analysis parameters
Number of other Individuals: 1
Invoice Amount Within: 1000
Within a timeframe of: 1 months ▼

[Start Risk Analysis](#)

Figure 23 Risk Analysis Page

Then, the user is redirected to the respective process page. In the Risk Analysis page, depending on the data type of the dataset, the user will have to select an appropriate risk analysis method. Then, depending on the risk analysis method selected, the user will have to set the appropriate attributes and parameters that pertain to the method. In a similar manner, in the Anonymization page the user will have to specify the attributes and parameters that pertain to an anonymization method as well as a .json file specifying the hierarchies of each attribute in the case that the selected method demands it. After the required fields are filled out, the user can initiate the process.

Viewing all initiated processes

After the initiation of a new risk analysis or anonymization process, the user is redirected to the Processing Queue page. In this page, previously initiated processes are displayed.

| Processing queue | | | | |
|------------------------------------|--------------------------|----------------------|-------------|-------------|
| Risk Analysis See All | | | | |
| Name ▾ | Status ▾ | Risk Analysis Method | Started | Ended |
| Tabular Dataset | Completed ● | k-anonymity | 8 days ago | 8 days ago |
| Spatiotemporal Dataset | Completed ● | Location | 1 month ago | 1 month ago |
| Textual Dataset | Completed ● | Textual | 1 month ago | 1 month ago |
| Tabular Dataset | Completed ● | l-diversity | 1 month ago | 1 month ago |
| Tabular Dataset | Completed ● | k-anonymity | 1 month ago | 1 month ago |

| Anonymization See All | | | | |
|------------------------------------|--------------------------|----------------------|---------------|------------|
| Name ▾ | Status ▾ | Anonymization Method | Started | Ended |
| Tabular Dataset | Running ● | k-anonymity | 2 seconds ago | - |
| Tabular Dataset | Completed ● | l-diversity | 2 days ago | 2 days ago |
| Tabular Dataset | Completed ● | k-anonymity | 2 days ago | 2 days ago |

Figure 24 Processing Queue Page

For each process in the queue, information about the status (“Running”, “Completed” or “Canceled”), the name of the Risk Analysis method used and the times of when the process started and when it ended are displayed. If a user wants to view all past processes, they can select the “See All” button beside each type of process (Risk Analysis or Anonymization) and a complete listing will be displayed.

Viewing the results of a process

A user can view the results of a process by:

- Selecting the last process of a specific dataset from the Datasets page
- Selecting a process from either the Risk Analysis or Anonymization pages
- Selecting a process from the Processing Queue page

After choosing a process, with one of the methods above, the user is either redirected to the appropriate results page in the case of a risk analysis process or has the ability to save the resulting dataset in the case of an anonymization process.

Viewing the results of a Risk Analysis process

After selecting a Risk Analysis process, the user is redirected to the De-anonymization Risk Analysis page.

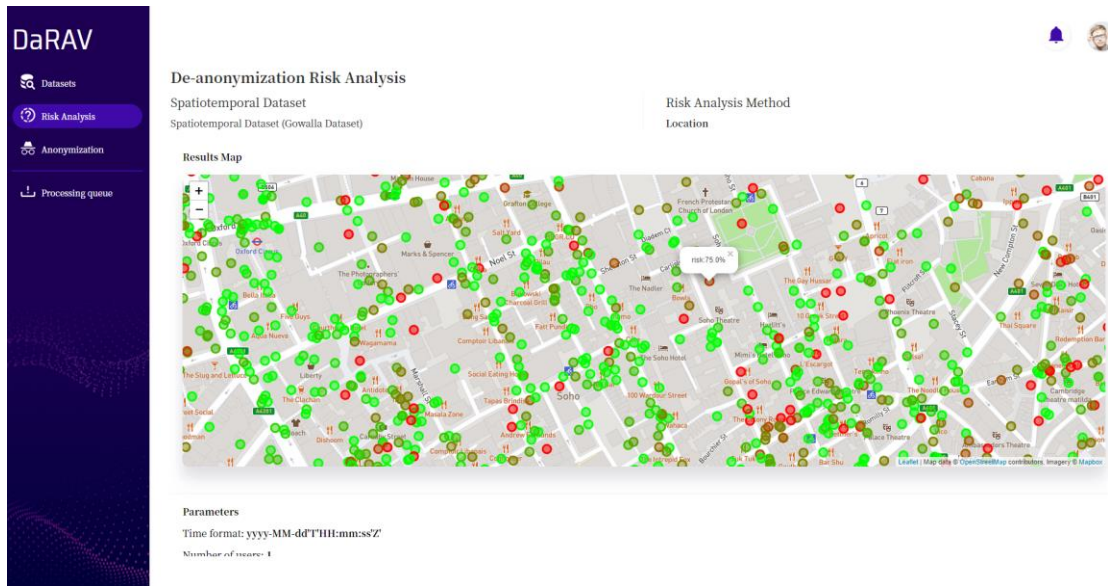


Figure 25 De-anonymization Risk Analysis Page

In this page, information regarding a specific risk analysis process is provided including metadata of the dataset under inspection. While the process' status is "Running" the user can only view the information related to the specific process and dataset. When the status changes to "Completed", the results of the process are displayed in the appropriate form (e.g. diagram or map) for the user to review them.

3.3 Analytics and Insights Application – InfoDrill

3.3.1 Application Description

InfoDrill is a Big Data visualization application designed and developed for enabling users to create targeted analyses on datasets. Through the application the users can correlate different datasets and analyze them by creating smart dashboards. Those dashboards can facilitate interactive data visualizations for data drilling as well as various complimentary visualization components (views). By interacting with a dashboard's drilling visualization, users can perform data drill down or roll up actions on the data, actions that will then propagate to the rest of the dashboard's views, allowing them to explore the data and discover insights.

About Datasets

In the context of the present application, the term “dataset” is used for the representation of a file in the user’s machine and for the combinations of already imported datasets. Specifically, the datasets that a user can have access to through the application are of two types:

- **Datasets**

The datasets that the user imports to the application.

- **Combined Datasets**

Datasets that are the result of the combination between two datasets that the user has already imported.

Each dataset has a set of metadata information that go along with it. Following are the data that need to be filled out by the user while importing a dataset in to the application:

- **File location and separator/delimiter**

The user can specify the file that they want to import from their local filesystem. The file has to be in .CSV format. They also have to set the type of delimiter (comma, semicolon or tab) used in the particular .CSV file.

- **Title and short Description**

The user can also set a title for the dataset as well as a short description for further explanation of the data contained in it.

The data and metadata related to the datasets imported to the application are all stored locally, in the user's machine which is running the application.

In order to create visualizations and review the analytics and insights that can be offered through the dashboards of the application the users have to first create a combination between two datasets. This can be achieved through the application's GUI. Users must choose two datasets as well as the attributes (columns) on which the correlation will occur. The two attributes must be of the same type and also belong to the types supported for correlation. The supported attributes for correlation are: geospatial, text, long, float, and date attributes.

Smart Dashboards

To enable our application to provide analytics and insights to its users, we chose to use the industry standard approach of allowing users to create dashboards and populate them with various data visualizations, called views. In addition to static data visualizations though, our smart dashboards provide users with the ability to drill down and roll up on the combination of two datasets, through the use of interactive visualizations. This "filtering" of data, which is the outcome of the users' drilling activities, is applied to all the views of the dashboard.

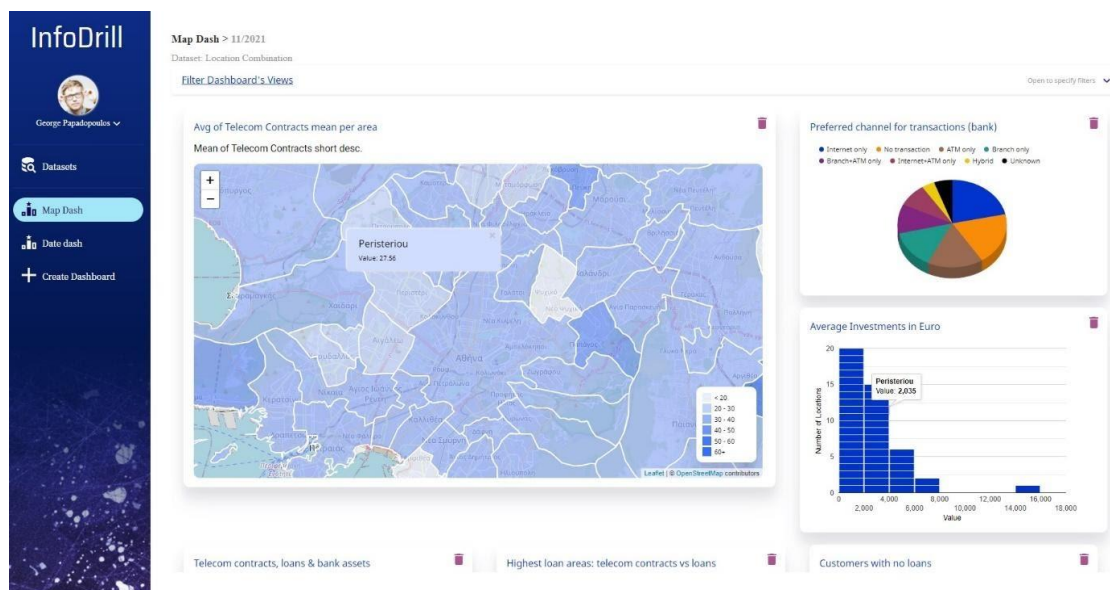


Figure 26 Smart Dashboard Example

In Figure 26 we can observe an indicative example of a smart dashboard that the users can create while using the application. In this example, a user has created a dashboard in order to analyze the combination of a financial and a telecommunications dataset. The data drilling visualization chosen for this dashboard is a choropleth map that the user has configured to display the average of telecom contracts per geographical area. In addition to that, the user has also created several views related to this analysis. One of them is a pie chart displaying a terms aggregation of the preferred channel for transactions of individuals of the financial dataset. Another view created by the user is a histogram displaying the average investments in euros, aggregated per geographical area.

As described in the beginning, if the user drills down to a specific location on the map, all of the dashboard's views will only query on data pertaining to the area shown. In this example, the user was initially provided with a map of the region of Attica. When the user zoomed in and focused the map above the general Athens area, the dashboard narrowed the data on which the views statistics are computed, from the whole Attica region to only the Athens area that was now focused on the map. In this manner, the user can see only the statistics for the Athens area in all the views and thus conduct a more focused analysis.

Data Drilling Visualizations

As mentioned in section 2.2.3, data drill down is the process in which a user can shift from a grouping of data to a more detailed and granular grouping of data of the same dataset while data roll up can be considered as the reverse process of data drill down. By interacting with the visualizations and filters offered by the dashboard the users can conduct data drill down and roll up processes that can offer more detail as well as narrow down the data being analyzed to the ones that the user is interested in.

Our application provides two visualization approaches for data drilling. The first approach is data drilling with the use of geospatial data visualized on geographical maps and the other with the use of temporal (date) data visualized on timelines.

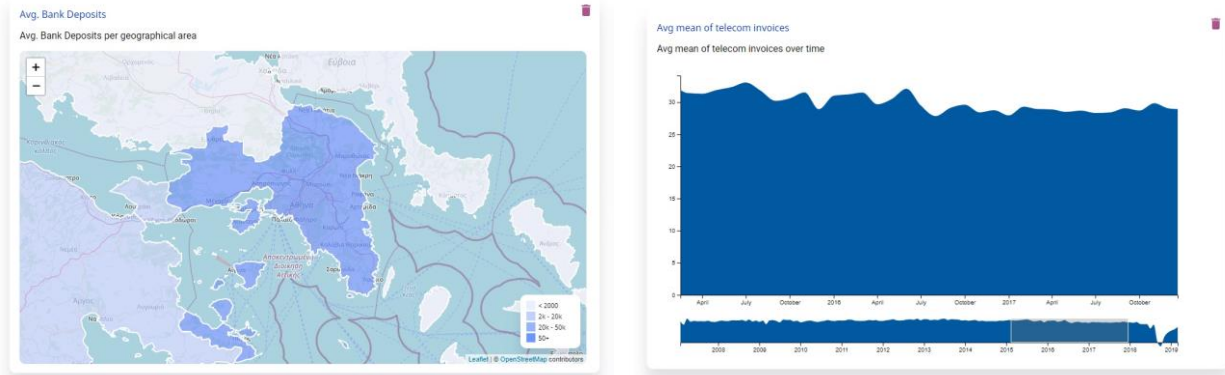


Figure 27 Data Drilling Visualizations, a geographical map (left image) and a histogram with a date brush (right image)

Specifically, the data drilling visualizations currently offered by the application are:

- **Choropleth Map**

A map in which geographical areas are colored based on the value of a spatially aggregated statistic of an attribute selected by the user. The coloring scheme is based on ranges for the aggregated value, also defined by the user. Depending on the map's zoom level, the map is displaying the regions or the municipalities of a country.

- **Heatmap Map**

A geographical map that depicts the intensity of the values of a spatially aggregated statistic of an attribute selected by the user. Areas of the map with higher intensity will be colored red, and areas of lower intensity blue.

- **Combination of Choropleth and Heatmap maps**

In this map configuration, the map switches from choropleth to heatmap in a predetermined zooming point. In this way users can evaluate statistics from the granularity of a country's regions, down to that of postal code areas-points.

- **Combination of Histogram with a Timeline**

This visualization combines a standard histogram plot with a timeline. This timeline provides the user with a "brush" (a gray area over the timeline, as can be seen on the right image of Figure 27). With the use of this brush, the user can zoom in or out on the histogram thus narrowing or widening the time frame of the analysis.

For each of the visualizations above the users will have to choose an attribute and then select which aggregate statistic of that attribute they want to visualize. Depending on the visualization type, this aggregation will be either spatial or temporal. The available aggregate statistics are: count, min, max, avg, sum, sum of squares, variance, population variance, sampling variance and std deviation. Depending on the geo-spatial or temporal nature of each data drilling visualization, the user will also have to specify an attribute of the dataset that is either of geospatial or date type in order to enable the visualization of the selected aggregate statistic.

Complementary Visualizations

In addition to data drilling visualizations, the dashboards provided by the application can also support visualizations that are complementary to the data drilling process. Although these visualizations cannot perform drilling actions on data directly, they do so indirectly by reducing their querying scope based on the data focused by the dashboard's data drilling visualization.

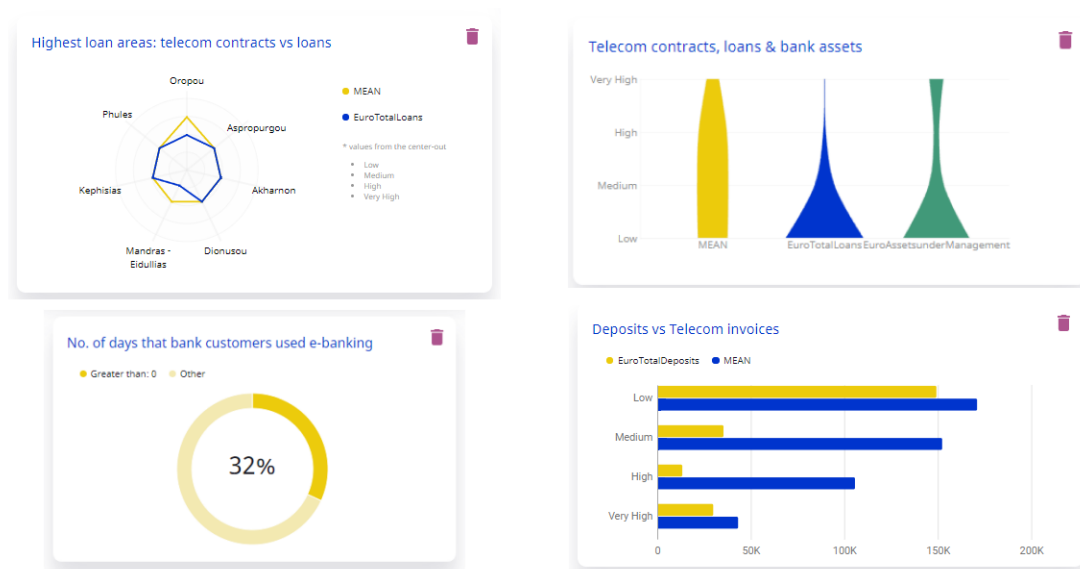


Figure 28 Examples of Complementary Visualizations, a radar chart (top-left), a violin chart (top-right), a donut chart (bottom-left) and a bar chart (bottom-right)

The visualizations currently provided by the application are the following:

- **Pie chart**

With this chart the users can either visualize a text type attribute's terms aggregation (meaning the distinct values that this attribute is composed of) or visualize a long or float type attribute's values in ranges.

- **Donut chart**

With this chart users can visualize the number of records for which an attribute of the dataset is equal to or not equal to a specified value. There is also the option to compare attribute values to specified values using greater or less than comparisons; however, this functionality is only available for long and float type attributes (see Figure 28).

- **Bar chart**

With this chart the users can visualize long or float type attributes values by creating custom ranges for each bar of the chart. This chart can facilitate up to two attributes being visualized at the same time, each one with separate, color coded, bars (see Figure 28).

- **Histogram**

With this chart the users can visualize an aggregated statistic of a float or long type attribute across all geographical areas where the data drilling visualization is currently focused.

- **Violin chart**

With this chart the users can see the distribution on a long or float type attribute's values. The users can visualize up to three attributes at the same time and they are also able to set custom buckets for the values of each one of the attributes (see Figure 28).

- **Radar chart**

With this chart the users can compare two long or float type attributes' values that would otherwise seem not comparable. For each attribute the user can set custom ranges for a specific aggregated statistic. In this way, a user is able to detect outliers or clusters of similar observations (see Figure 28).

The available aggregate statistics are: count, min, max, avg, sum, sum of squares, variance, population variance, sampling variance and std deviation.

3.3.2 Implementation

Application Architecture

Respectively to DaRAV, this application also followed the same architecture approach, according to which every component constitutes a standalone entity, realized in a separate docker image. All the components/images needed for the application can be deployed through a single docker-compose file.

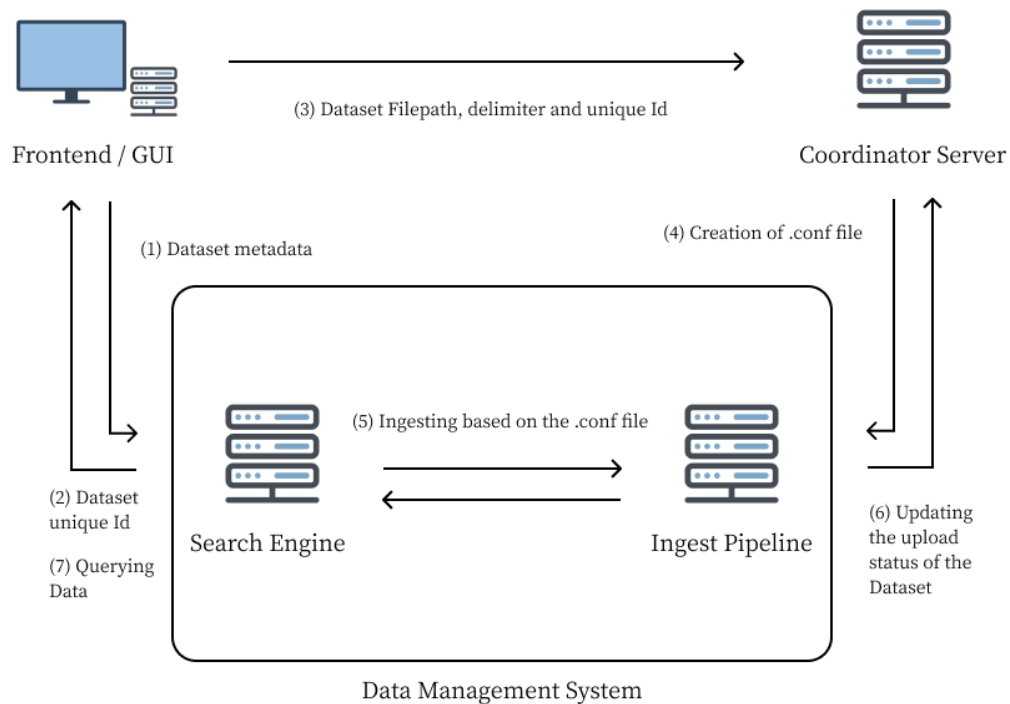


Figure 29 InfoDrill Application Architecture diagram

The main components of the app's architecture are the following:

- The application frontend/GUI
- The Coordinator Server, which is responsible for the necessary backend operations with regard to the coordination of the backend components according to specific workflows (e.g., data ingestion to the Data Management System from files, updating the upload statuses, etc.).

- The Data Management System, comprising the Search Engine and Ingest Pipeline. It is responsible for the communication between the components (metadata), storing of the datasets destined for analysis as well as the logging of the App's components outputs.

In Figure 29, the diagram of the architecture depicts the flow of information between the API's of the individual components. These workflows are mainly triggered by the user's actions (as it will be described in the Using the Application section). Following is a brief description of each one:

1. Dataset Metadata

The user chooses a dataset to be "uploaded" (imported) and enters metadata for it, similarly to the DaRAV's dataset import process, by filling out the upload page's form.

2. Dataset unique Id

Through an AJAX Request the frontend/GUI sends these metadata to the Data Management System, where a new document (that contains them) is created in the index responsible for storing the metadata related to datasets.

3. Dataset File path, delimiter and unique Id

The unique Id of the newly created document (containing the dataset's metadata) as well as the dataset's file path and delimiter are sent to the Coordinator Server through its API.

4. Creation of .conf file

Based on the metadata received, the Coordinator Server creates a configuration (.conf) file. This file is needed in order for the Ingest Pipeline to create a new pipeline that will ingest the dataset to the Search Engine.

5. Ingesting based on the .conf file

When the Ingest Pipeline detects the new configuration file, it sets a new pipeline, as per the parameters specified in the configuration file. This pipeline ingests the .CSV file to the Search Engine by auto detecting the types of fields (textual, numeric, date) as well as mapping geospatial fields to location identifiers associated with postal codes, municipalities and regions based on mapping schemes specified by the .conf file.

6. Updating the upload status of the Dataset

After the start of the new pipeline, the Coordinator Server monitors the ingesting progress and updates the dataset’s metadata. After the ingestion process is completed, the dataset is ready for use.

7. Querying Data

When the user (through the GUI) combines two already uploaded datasets, a “Combined Dataset” is created and is ready for analysis. At this point, the Frontend / GUI starts to query the indices of the two Datasets (that the “Combined Dataset” is composed of) for analysis purposes.

Smart Dashboards and Dataset Correlation

The smart dashboards offered by the application, as stated in previous sections, can support simultaneous data drilling on both datasets that the combined dataset consists of.

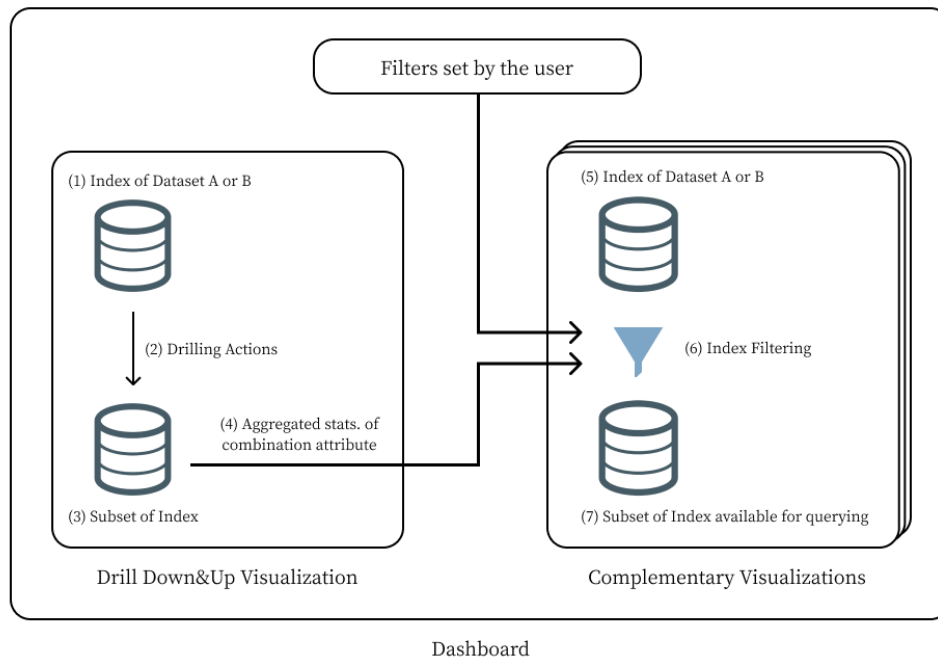


Figure 30 Smart Dashboard implementation diagram

This posed some interesting technical challenges on how it could be implemented. In our approach, we decided to implement this correlation by filtering the dataset indices on the fields (dataset attributes) that would be chosen from the users during the formulation of the combined dataset. In Figure 30, the diagram of the Smart Dashboard implementation depicts the flow of information between the dashboard’s visualizations. Following is a brief description of each one:

1. Index of Dataset A or B

For each of the datasets, named “A” and “B”, of the selected combined dataset there is a corresponding index in the Search Engine of the application containing its data. During the set-up of the dashboard’s data drilling visualization, the user has to specify which fields of either the index of dataset A or B will be utilized for the visualization.

2. Drilling Actions

Initially, the visualization offers the user an aggregated view of the selected fields for all the documents of the index. After the user’s interaction with the visualization, the documents that are taken into consideration for computing the statistics are narrowed down to a subset of the index. This is achieved through the use of geo queries (queries for finding the geographical areas within the map’s current bounding box) as well as date aggregation queries.

3. Subset of Index

This is the subset of the index that is being visualized after each interaction of the user with the data drilling visualization.

4. Aggregated statistics of combination attribute

From the subset of the dataset index that is visualized by the data drilling visualization, aggregated information for its’ selected attribute for combination is extracted. Depending on the data type of the attribute selected for combination, different aggregated information needs to be acquired from the subset of the index:

- **Long of Float types**

If the type of the attribute is of numeric nature, then the minimum and maximum values of the field are being acquired.

- **Date type**

If the type is date, then the oldest and newest date values of the field are being acquired.

- **Geospatial type**

If the type of the attribute is geospatial, then the unique geographical area ids for the field are being acquired.

- **Text type**

If the type is text, then the unique string values of the field are being acquired.

5. Index of dataset A or B

Just like the data drilling visualization, in order for users to set up a complementary visualization they have to select a field of either the dataset index A or B that they want to visualize.

6. Index filtering

Before executing the queries for acquiring the statistics required to visualize the field, the index to which the field belongs to must be filtered. The values of the attribute for combination that were retrieved in step four (4) are set as conditions for the documents that will be included in the subset of the index about to be visualized.

Specifically, if the combination attributes are of numeric or date type, the documents included on the indices are those whose values fall within the minimum and maximum values retrieved in step four (4). In the case of geospatial or text attributes, the documents included are those whose values belong to the set of unique values derived in step (4). As a result, all dashboard visualizations query their data from subsets of either index A or B, which contain documents that have the same ranges of values for the combination attributes.

Additionally, if the users wish to set their own filters for some attributes of either datasets, they can do so through the application's GUI. These filters will also be applied on this step.

7. Subset of index available for querying

This is the subset of the index that is being visualized by the complementary visualization.

Technologies used

One of the most challenging factors for developing this application was that of data storage and querying. Because of the potentially large size of the datasets (millions of records and gigabytes of data), the usage of a database was mandatory. Furthermore, given the size of those datasets and the requirement for our application to provide combinations between them, using SQL JOIN or set procedures (e.g., UNION, INTERSECT, etc.) could be both time consuming and

computationally demanding. Additionally, due to the unpredictability of the structure and types of datasets imported by users, the need for dynamic schemas to support unstructured data arose. Finally, advanced querying, such as geo querying, was also required for the analysis of such datasets. Based on these considerations, we determined that using a non-relational database would be more advantageous than using a relational database in terms of meeting the application's requirements.

Over a range of such data storing solutions, we decided to implement the Data Management System of the application based on the Elastic Stack. Similarly to DaRAV's utilization of Elasticsearch [32], InfoDrill utilizes Elasticsearch for the storing, querying as well as the communication requirements of the application. Furthermore, InfoDrill makes use of search engine's extensive REST Search API which ranges from aggregation to geo queries that can search data stored in one or multiple indices. Logstash [43], in the context of this application, is used for creating pipelines that can ingest and transform the data that users want to analyze and then stores them to Elasticsearch indices.

For the frontend / Graphical User Interface (GUI) of the application, the framework used was Angular v10 [36]. For the geographical maps, charts and graphs offered by the application, various libraries such as D3.js [44], leaflet [37] and Google Charts [45] were used.

For the Coordinator Server Node.js was used. Node.js is an open-source, cross-platform, JavaScript runtime environment, built on Chrome's V8 JavaScript engine [46]. Specifically, Node.js was used as an intermediate server (with the use of the back-end, web application framework Express.js) for the communication between the Angular framework, Logstash and the Elasticsearch node.

3.3.3 Using the Application

After the successful deployment of the application, users can have access to the Graphical User Interface (GUI) by visiting the specified link on an internet browser. Upon entry, the users will be asked for their credentials in order to log in. After a successful authentication, the users are able to view all the imported datasets as well as the combined datasets that resulted from previous combinations of the already imported datasets. Users can create new dashboards for each combined dataset by filling out a form with all of the required information, such as the configuration of the datasets' attribute statistics displayed by the main drilling visualization. Users can also add complementary visualizations to each dashboard based on the needs of their analysis. As a result, through these dashboards, users can focus their attention on aspects of the datasets where there is a high likelihood of uncovering unusual information and thus arriving at new, previously undiscovered insights.

User Login and Application Layout

The users can log in to the application by entering their credentials. If the authentication process succeeds, they proceed to the Datasets Page. After a successful authentication, they are directed to the Datasets page. All the pages offered by the application follow a similar design pattern to that of DaRAV (see Figure 31).

From the horizontal menu, users can navigate through the main pages of the application as well as view their profile which is located at the top of the sidebar. The main navigation option of the menu is the “Datasets” button that directs to the Datasets page. After choosing any of the combined datasets, the sidebar displays navigation buttons for the available dashboards of selected dataset as well as the “Create Dashboard” button for directing the users to the page where they can create new ones (see Figure 34).

Uploading a new Dataset

Users can import a new dataset by selecting the “Upload Dataset” button located at the top-left of the Datasets page (see Figure 32).

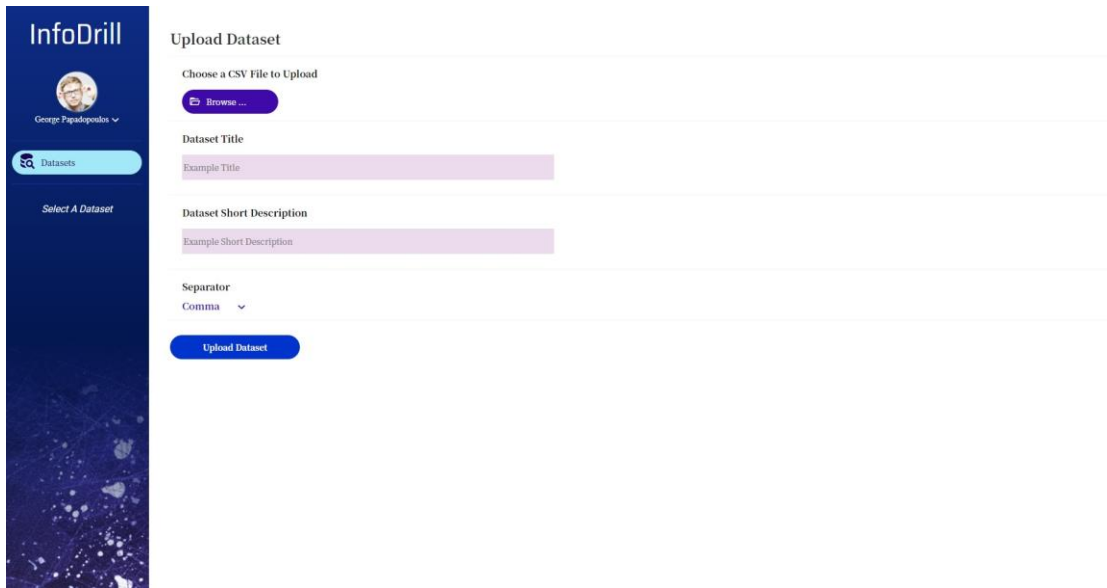


Figure 31 Upload Dataset Page

In the Upload Dataset Page, a user can specify the .CSV file containing the dataset, the dataset's title, short description and delimiter/separator of the .CSV file (see Figure 31). After filling out this information, the importing process begins. Until this process is finished, the user is only able to edit or delete the dataset. After the completion of the process, the dataset is ready to be combined.

Viewing the Available Datasets

In the Datasets page, users can view all the datasets imported to the application as well as the datasets that were created from their combination. There are two viewing options: the card view (set by default) and the list view. The user can specify the desired view with the use of the buttons located at the top-left of the page (see Figure 32). The datasets are organized in tabs based on their type.

There are two types of datasets:

- **Datasets**

The datasets that the users imported to the application

- **Combined Datasets**

Datasets that are the result of the combination between two datasets that the users have already imported

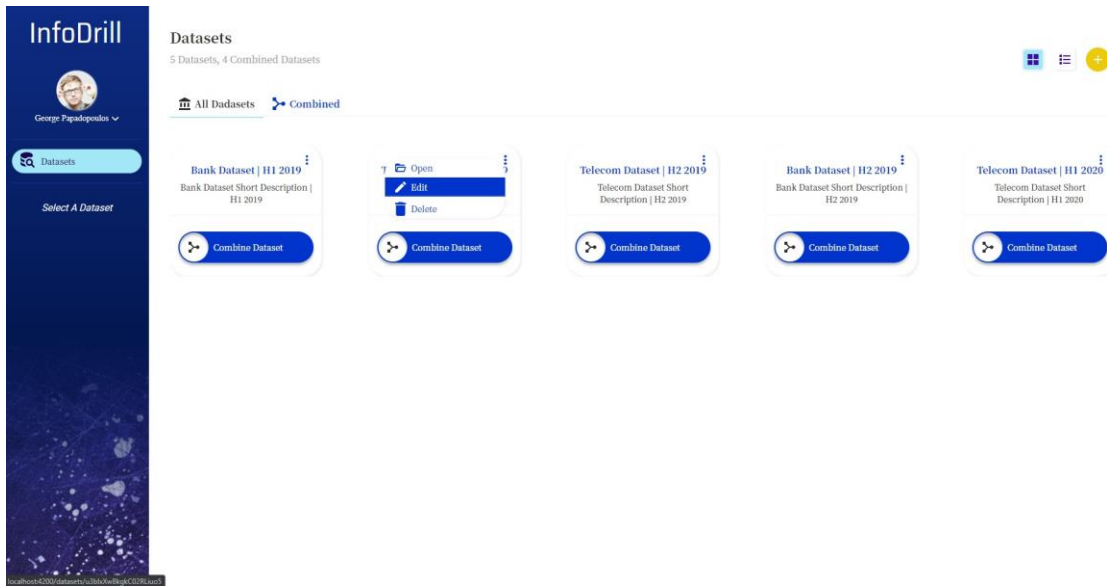


Figure 32 Datasets Page – all datasets tab – card view

For each dataset, the title, description and the ability to combine it with another dataset are displayed. From the options menu (three blue dots) the user has the ability to edit or delete the dataset. While a dataset is being uploaded, the user is only able to edit or delete it. After the uploading process is completed, the ability to combine the dataset with another is enabled.

Combining Datasets

A combination of two datasets need to be formulated in order for users to be able to create and

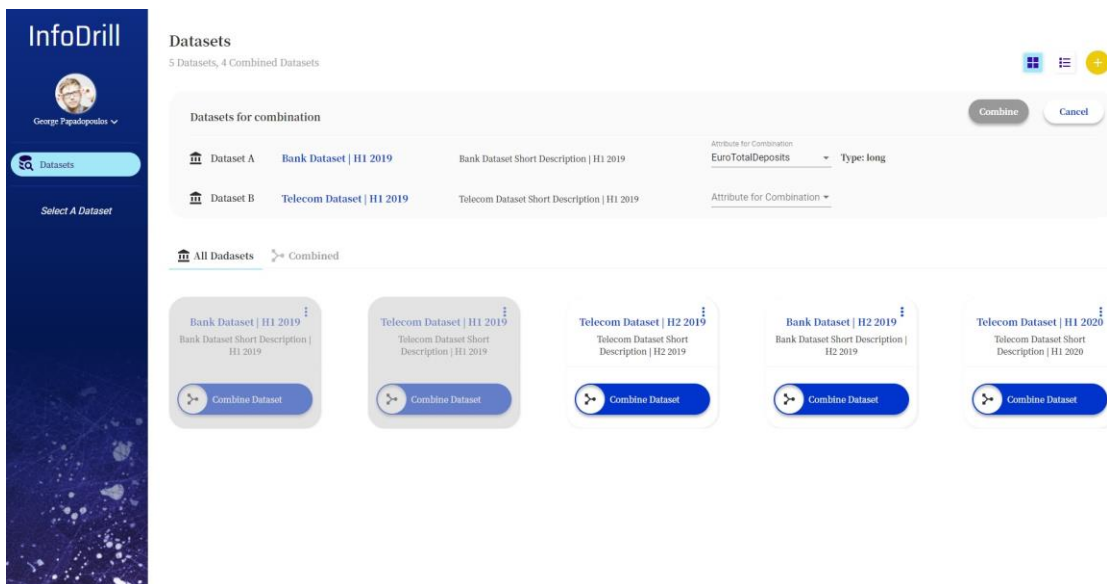


Figure 33 Datasets Page – Combining Datasets

review dashboards. Selecting the "Combine Dataset" button on a dataset will add the dataset in the list of datasets about to be combined. After selecting two datasets to be combined, users must select which attributes (columns) of the datasets they want this correlation to occur on. The users can select any attribute from each dataset, provided that its data type is one of location, text, long, float, or date. The selected attributes of the datasets must be of the same data type. When all of the preceding requirements are fulfilled, the users can proceed with the combination by selecting the "Combine" button (see Figure 33).

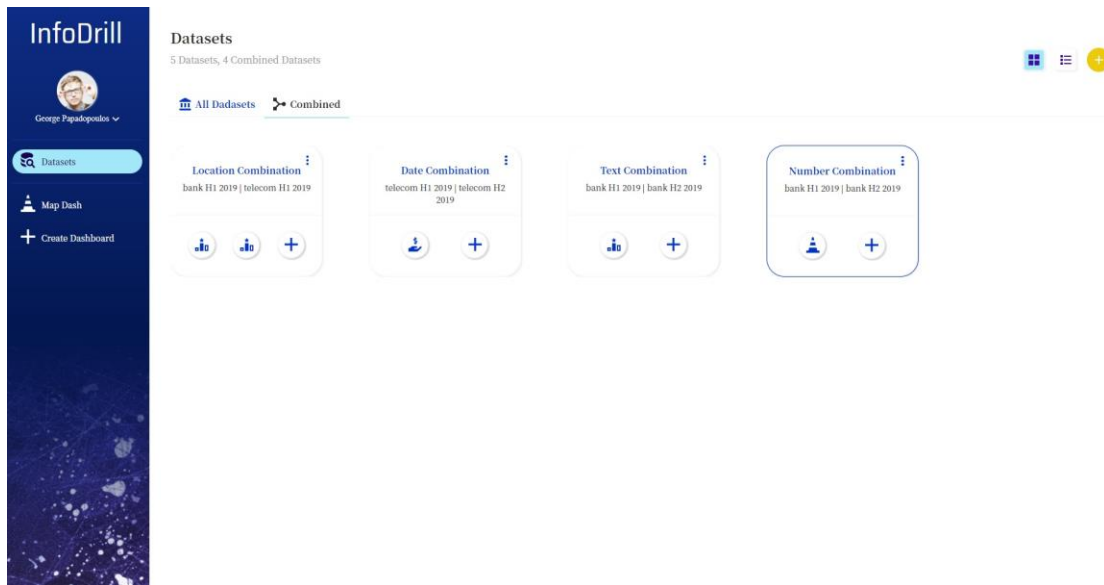


Figure 34 Datasets Page – Combined Datasets

Following the completion of the combining process, a new combined dataset is created and added to the Combined datasets tab (see Figure 34). From this dataset, users can create dashboards that are based on the correlation of the datasets that compose it. Users can also quickly select one of the recently created dashboards from the dataset's card, or choose one from the sidebar's full listing of previously created dashboards.

Editing Datasets

After selecting the "Edit" option of a dataset, the user is directed to the dataset's info page. In this page, all the basic information of the dataset is displayed. The user has the ability to edit the dataset's title and short description by selecting the pen icon beside either its title or short description (see Figure 35).

InfoDrill

George Papadopoulos

Datasets

Map Dash

Create Dashboard

Dataset Info

Dataset Title

Location Combination

Dataset Short Description

bank H1 2019 | telecom H1 2019

Dataset Type

combined

Dataset A Id

mX7oxXwBkgkC02RLSTGO

Dataset B Id

u3blxXwBkgkC02RLlao5

Dataset A Combination Attribute

NewPostalCodeID

Dataset B Combination Attribute

INITIATION_DEALER_ID

Update Dataset

Figure 35 Dataset Info Page

Create a new Dashboard

In order for the users to create and review analytics for a combined dataset they have to create a dashboard. To create a dashboard, the users have to select either the “Create Dataset” button from the card of a dataset or select it and then choose the “Create Dataset” button from the sidebar.

InfoDrill

George Papadopoulos

Datasets

Digital Services

Create Dashboard

Create Dashboard

Dashboard Title

test

Dashboard Icon

Attribute for Data Drilling

Attribute for Drilling: EuroLoans Type: long

Drilling Visualization

Choropleth Map

Heatmap Map

Choropleth + Heatmap

Histogram + Date

Select Geospatial Attribute

Select Attribute

Visualization Title

Example Title

Figure 36 Create Dashboard page

On the Create Dashboard page, the users are presented with a form where various aspects of the dashboard need to be specified. These are the following:

- **Dashboard Title**
The title of the dashboard that the users will be able to identify it with.
- **Dashboard Icon**
An icon that will precede the title of the dashboard.
- **Attribute for Data Drilling**
The attribute of either dataset A or B which will be analyzed from the Data Drilling visualization.
- **Drilling Visualization**
The visualization the users wish to use for data drilling on the chosen attribute. The users can select only a visualization compatible with the type of the attribute they selected. The visualizations that are not compatible are greyed out.
- **Geospatial or Date Attribute for the Drilling Visualization**
Depending on the visualization chosen, the application will require an attribute to base the visualization on.
- **Visualization title**
The title of the visualization, as it will be displayed on the dashboard.
- **Visualization short description**
A short description of the visualization, as it will be displayed on the dashboard.
- **Attribute's aggregated statistic for comparison**
The aggregated statistic the users wish to visualize through the selected visualization. There is a great variety of aggregated statistics for the users to choose from, e.g.: count, min, max, avg, etc.
- **Aggregated statistic ranges for choropleth map**
Here the users must specify the ranges of the aggregated statistic by setting the “from” and “to” values for each one as well as setting a name for the legend of each range. These are values that are required only for the choropleth map visualization.
- **Maximum value of statistic**

This is the maximum value that the users want to limit the results that correspond to the selected attribute's statistic. This is a value required only for the heatmap map visualization since based on that value, the heatmap can calculate the magnitude coloring.

After the users have filled in the entire form they are able to create the dashboard by selecting the “Create Dashboard” button. Then, they are directed to the page of the newly created dashboard.

Dashboard Page

The analytics provided by this application are all delivered via dashboards that include visualizations that the users have set up and customized. The design of all dashboard pages follows the same pattern as the one that can be seen on Figure 37.

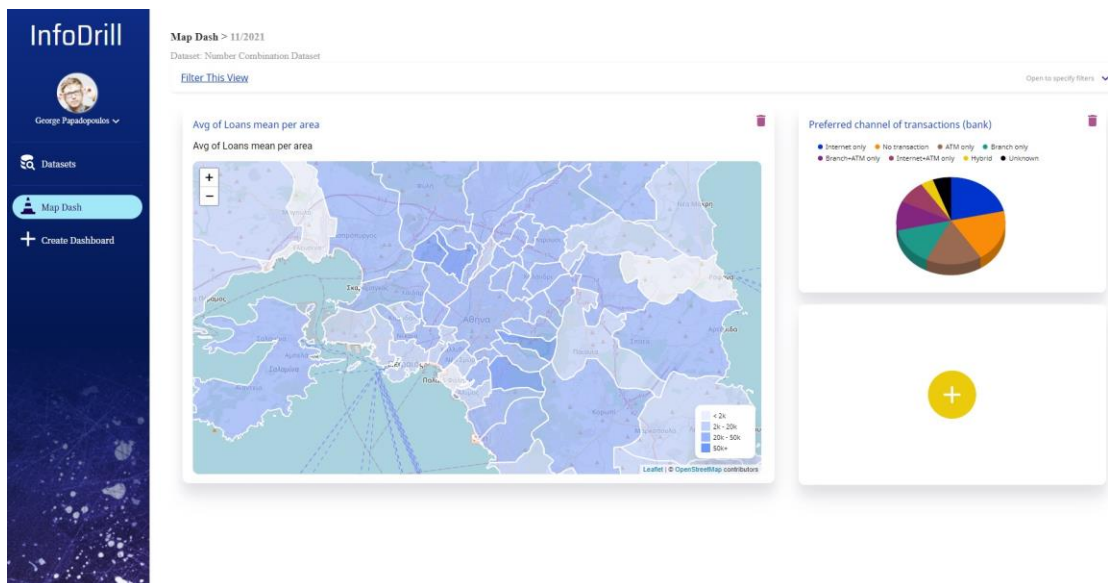


Figure 37 Dashboard page

The title of the dashboard, as well as the date it was created, are displayed from top to bottom, followed by the name of the combined dataset for which this dashboard was created. Following that, there is the Filters dropdown menu, where users can select filters that are applied to all dashboard visualizations. Then there's the section where the drilling visualization as well as all of the complementary ones lie. The Drilling Visualization is at the top left of this section, with the complementary visualizations placed around it. After creating a dashboard, the only existing view

is the drilling visualization. The users can add views by selecting the yellow plus button on one of the five available view positions. Each dashboard can have a total of six views, the data-drilling one and five complementary. The users have also the option to delete a view, and thus free a position. By deleting the drilling visualization, the dashboard is also being deleted.

Create a new View

Users can create a new view by selecting the yellow plus button of an available position on a dashboard. Then, they will be directed to the Create View page.

The screenshot shows the 'Create View' interface. On the left is a sidebar with the 'InfoDrill' logo and user information for 'George Papadopoulos'. The main area is titled 'Create View' and includes a 'View Title' field with the text 'test'. Below this is the 'Attribute(s) for Visualization' section, which has a dropdown menu set to 'EuroDeposits' and 'Type: long', along with an 'Add Attribute' button. The 'View Visualization' section displays six chart type options: Pie Chart, Donut Chart, Histogram, Radar Chart, Violin Chart, and Bar Chart. At the bottom, there is a 'Statistics Calculation' section.

Figure 38 Create View page

On the Create View page, the users are presented with a form where various aspects of a view need to be specified. These are the following:

- **View title**

The title of the view, as it will be displayed on the dashboard.

- **Attribute(s) for Visualization**

The attribute(s) of either dataset A or B which will be analyzed. The user can choose to analyze from one (1) and up to three (3) attributes at the same time.

- **View Visualization**

The visualizations that are available for this view. Depending on the number and data type of the selected attributes for visualizing, the available visualizations may change. The users can select only a visualization compatible with the types of the attributes they selected. The visualizations that are not compatible are grayed out.

- **Statistics Calculation**

The calculations on which the view will be based on. Depending on the type of visualization selected the calculation options may change.

After the users have filled in the entire form they are able to create the view by selecting the “Create View” button. Then, they are directed to the page of the dashboard where the newly created view is being displayed.

Filters

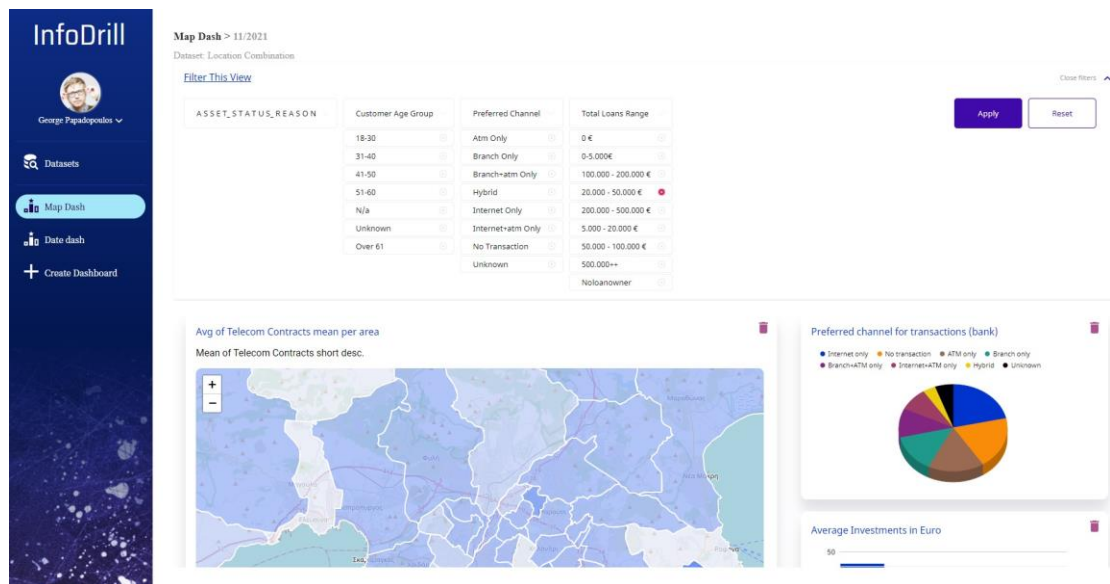


Figure 39 Dashboard page – Filters

If the aforementioned exploration isn't sufficient enough for the users' needs, they are also able to apply filters to the dashboard's views that they are currently reviewing by selecting the “Filter Dashboard's Views” link on the top left of every dashboard page. There, they can view the unique values of the fields that are included in the datasets that are being analyzed. By selecting to remove a value from a specific field they are also limiting the records that are taken into consideration by the Application in order to display the Analytics of the current page. The users

can reset all the values that were removed by selecting the “Reset” button (see Figure 39). The filtering function is scoped to the dashboard currently being viewed, meaning that if the users move to a different dashboard, the filters reset automatically.

Chapter 4

Evaluation

In this section, we will discuss our approach at conducting the framework's evaluation study. This study is composed of two parts: a heuristic evaluation for the deanonymization - risk analysis application, DaRAV, and a user testing evaluation for the analytics and Insights application, InfoDrill. We will also review the results of these evaluations as well as discuss their findings.

4.1 Methodology

4.1.1 Heuristic Evaluation of DaRAV Application

The DaRAV's heuristic evaluation was carried out by UX experts following an evaluation approach offered by J. Nielsen [47]. This approach is based on the following heuristic evaluation guidelines:

1. **Visibility of system status:** The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2. **Match between system and the real world:** The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
3. **User control and freedom:** Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
4. **Consistency and standards:** Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5. **Error prevention:** Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
6. **Recognition rather than recall:** Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7. **Flexibility and efficiency of use:** Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
8. **Aesthetic and minimalist design:** Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
9. **Help users recognize, diagnose, and recover from errors:** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
10. **Help and documentation:** Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Each of the evaluators independently inspected the application's GUI based on these guidelines, and each problem that was identified was associated with one or more guidelines. The evaluators assigned a severity rating to each identified issue. These ratings ranged from zero (0) to four (4) and were based on the frequency, impact, and persistence of the problem, as follows:

- **0:** I don't agree that this is a usability problem at all
- **1 - Cosmetic problem only:** need not be fixed unless extra time is available on project
- **2 - Minor usability problem:** fixing this should be given low priority

- **3 - Major usability problem:** important to fix, so should be given high priority
- **4 - Usability catastrophe:** imperative to fix this before product can be released

4.1.2 User Testing Evaluation of InfoDrill Application

For the user testing of InfoDrill, the evaluation mainly focused on the application's usability, user experience, and overall performance, as well as the workload imposed by the application to users. The testing of the application was conducted with remote testing sessions through the use of a teleconference application. The users taking part in these sessions were asked to follow specific usage scenarios that were provided by the evaluator responsible for the session. These scenarios were designed to test the application's key functionalities. During their execution by the user, the session's evaluator was responsible for observing and taking notes of the user's actions. The study and its protocol were approved by the Ethics Committee of FORTH-ICS (Reference Number: 133/17-11-2021).

The first scenario (as seen in Table 1) included tasks aimed at evaluating the dashboards that can be created from the application as well as the statistical information they can provide to the users. The data exploration functionalities of the dashboards were also tested, as well as whether the results of these explorations can help users perceive and derive insights for the data being analyzed.

Table 1 Scenario A: Dashboard Comprehension

| Task | Description |
|--------|--|
| Task 1 | Locate an existing dataset combination and open an already created dashboard. |
| Task 2 | After opening the dashboard from Task 1, locate on the map a municipality with a high value of the visualized statistic and select it. |
| Task 3 | Identify specific statistics from the dashboard's views and whether those statistics can be correlated with the high value of this municipality. |
| Task 4 | Locate on the map and select a municipality with low value of the visualized statistic. |

CHAPTER 4. EVALUATION

| | |
|--------|--|
| Task 5 | Identify specific statistics from the dashboard's views and whether those statistics can be correlated with the low value of this municipality. |
| Task 6 | Is there a connection between the values of the municipality and the statistics observed? Can an insight be derived? |
| Task 7 | Locate a specific municipality on the map and judge if this municipality complies with the insight derived from Task 6. Should specific promotional actions be taken on this municipality? |

The second scenario (as seen in Table 2) consisted of tasks designed to evaluate the procedure for combining two datasets in order to generate a new combined dataset from previously uploaded datasets. This scenario also evaluated the creation of a new dashboard and the configuration of its data drilling visualization, as well as the addition and configuration of views to that dashboard.

Table 2 Scenario B: Dashboard Creation

| Task | Description |
|----------|--|
| Task 8.1 | Create a new dataset combination from two already uploaded datasets. Combine these datasets on a specific attribute. |
| Task 8.2 | Rename the newly created combined dataset with a given title. |
| Task 9.1 | For the combined dataset created at Task 8.1, create a new dashboard and name it with a given title. |
| Task 9.2 | For the data drilling visualization of the dashboard, visualize a specific attribute per geographical area, with the use of a choropleth map. |
| Task 9.3 | Set the aggregate statistic per geographical area for that attribute to be the average of the values. Set ranges for this statistic based on the given range values. |

| | |
|-----------|--|
| Task 10.1 | For the dashboard created at Task 9, add a new view that visualizes a specified attribute. This visualization has to be a donut chart. |
| Task 10.2 | Set the statistic comparison of the view based on a given comparison and value. |

In each session, the evaluator kept notes of the user's success at completing tasks of each scenario, the errors that the user made at each task, and the amount of help given to the user after their request. For measuring the workload the users experienced during the execution of the scenario's tasks as well as their overall experience, the evaluator offered to the users' an online questionnaire for them to fill up. This questionnaire was based on NASA's Task Load Index (NASA-TLX) as well as the UMUX-Lite.

The NASA's Task Load Index (NASA-TLX) is a multidimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales [48]:

- **Mental Demands:** How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving?
- **Physical Demands:** How much physical activity was required? Was the task easy or demanding, slack or strenuous, restful or laborious?
- **Temporal Demands:** How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
- **Performance:** How successful do you think you were in performing the task? How satisfied were you with your performance in accomplishing this task?
- **Effort:** How hard did you have to work (mentally and physically) to accomplish your level of performance?
- **Frustration Level:** How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

The NASA-TLX questionnaire consists of the aforementioned questions. The users can respond to these questions by selecting a value between zero (0), which indicates a low endpoint, and twenty (20), which indicates a high endpoint for each of the subscales. The only exception is the

Performance subscale, where a score of zero (0) indicates good performance and a score of twenty (20) indicates poor performance. The Task Load Index can be computed by calculating the average of the ratings of the questionnaire.

The questionnaire that the users were given included, in addition to the NASA-TLX questionnaire section, a UMUX-Lite section. Usability Metric for User Experience (UMUX) is a metric intended for assessing the perception of the ease of using of a system by a user [49]. It contains two positive and two negative constructs that can be answered with a 7-point response scale. UMUX-Lite [50] is a shorter version of the UMUX questionnaire where the users have to respond to two constructs:

- This system's capabilities meet my requirements
- This system is easy to use

The users can respond to these constructs with a number from one (1) to seven (7), where one (1) indicates that the user totally disagrees and 7 indicates that the user totally agrees with the construct's statement.

At the end of each remote evaluation session, the users had a debriefing interview with the evaluator. Specifically, they had to answer the following brief questions:

- What did you like more about the application?
- What would you like to be changed or improved in the application?
- Have you used a similar application in the past?

These questions aimed at capturing the users' thoughts about the system and their overall experience as well as their familiarity with such systems as InfoDrill.

4.2 DaRAV Heuristic Evaluation

4.2.1 Procedure

The DaRAV application was heuristically evaluated by three UX Experts, each of whom assessed the application independently. Each expert's problems were collected, and the data was then aggregated into a single report, removing duplicates and merging identical problems. Following that, each evaluator provided their severity rating for each of the problems in the unified list, with a final severity rating calculated as the mean of the individual evaluators' ratings.

4.2.2 Results

The heuristic evaluation of the DaRAV application resulted in a final evaluation report with 23 issues, 9 of which were minor or cosmetic in nature (severity ≤ 2) and 5 of major severity (severity ≥ 3), amongst others. The complete listing of the issues discovered during this evaluation can be seen on Table 3.

Table 3 DaRAV Heuristic Evaluation Results

| Screen / Action | Issue Description | Guidelines | Severity |
|------------------|--|------------|----------|
| General | The aims and objectives of this application are not clear at first. Perhaps a short text passage in the login screen would help to resolve this issue. | (6), (10) | 1.67 |
| Datasets Screen | Tooltips should be presented on top of interactive buttons to assist first-time users in understanding their functionality. | (6), (10) | 2.67 |
| Datasets Screen | The delete dataset icon should be smaller. | (8) | 2.00 |
| Selected Dataset | Upon selecting a dataset there is not a back button. | (3) | 2.33 |
| Selected Dataset | It should also be possible to update the .CSV file on the Dataset Info page. | (5) | 3.33 |

CHAPTER 4. EVALUATION

| | | | |
|----------------------|--|----------------|------|
| Selected Dataset | The uploaded .CSV file's value should not be bold because it conveys the meaning that it is a category. | (8) | 1.00 |
| Selected Dataset | The value of the dataset title should not be blue (it is not a link). | (8) | 1.67 |
| Selected Dataset | The value of the separator should not be bold, since it conveys the meaning that it is a category. | (8) | 1.00 |
| Selected Dataset | From here, direct action buttons for risk analysis and anonymization should be provided. | (7) | 2.67 |
| Risk analysis Screen | In the "Risk Analysis Method", there is no meaning in having a drop-down when there is only one single option. | (7) | 1.33 |
| Risk analysis Screen | There are no instructions at all on this screen. They would have been very useful to assist users understand what they have to select. | (5), (6), (10) | 2.83 |
| Risk analysis Screen | The option buttons for each of the datasets' columns should only select one option. | (5) | 2.67 |
| Risk analysis Screen | Processing parameters and risk analysis parameters are not well-aligned. | (8) | 1.33 |
| Risk analysis Screen | Processing parameters are impossible to be filled-in correctly (e.g. time format expects text input, with no guidance at all). | (5), (6) | 3.17 |
| Risk analysis Screen | Risk analysis parameters should also provide guidance and restrict users to values that they can enter. | (4), (5), (7) | 3.00 |
| Risk analysis Screen | Errors are not helpful ("Error in form"). | (9) | 3.67 |

CHAPTER 4. EVALUATION

| | | | |
|------------------------------|---|----------------|------|
| Anonymization Screen | In the "Risk Analysis Method", there is no meaning in having a drop-down when there is only one single option. | (7) | 1.33 |
| Anonymization Screen | There are no instructions at all on this screen. They would have been very useful to assist users understand what do they have to select. | (5), (6), (10) | 2.83 |
| Anonymization Screen | The option buttons for each of the datasets' columns should only select one option. | (5) | 2.67 |
| Anonymization Screen | Parameters and options at the bottom of the list are not well aligned. | (8) | 1.33 |
| Anonymization Screen | Errors are not helpful ("Error in form"). | (9) | 3.67 |
| Risk analysis results Screen | The analyst should be able to directly exclude points from the dataset, re-run risk analysis or start an anonymization process, otherwise revert to the original dataset. | (3), (7) | 2.83 |

In summary, the most important problems were about the forms that configure a risk-analysis or anonymization procedure. All the issues that have been identified by the heuristic evaluation will be addressed in future DaRAV versions.

4.3 InfoDrill User Testing Evaluation

4.3.1 Procedure

For the user testing evaluation of the InfoDrill application, the users participating were 20 executive, marketing, data analysis and technical personnel from telecommunications and financial companies as well as local municipalities. They were contacted via email and were given all the necessary information in order to take part in the remote evaluation session.

At the start of the session, prior to the test, the users were given a consent form that they had to fill up before participating. This consent form informed the users about the study that they were going to participate in, the benefits of their participation to the improvement of the application as well as the handling of their personal data.

Before handing the users tasks to execute in the application, the session evaluator provided a brief system walkthrough explaining the main premise of the application as well as some of its core functionalities. The users were then instructed to open the application on their browser using a link provided by the evaluator. They then shared their browser window, so that the evaluator could observe their actions during the testing.

During the testing phase of the evaluation, the users were asked to execute the tasks of the two scenarios. The evaluator offered the tasks one by one, both orally and in text form (with the use of the teleconference application's chat). While the task was executed by the users, the evaluator was available for help if the users requested it or if they were stuck for a significant time. The evaluator, at the same time, was keeping notes about the success of each task, the help given and the errors users made.

After the completion of each scenario, the users were asked to fill up a NASA-TLX questionnaire for the workload that they had experienced at that point. With the completion of both scenarios, the users were then asked to fill up a UMUX-Lite questionnaire in order to evaluate the overall user experience. Each session concluded with a debriefing interview where the users were asked a few brief questions from the evaluator regarding their experience with the system as well as their familiarity with systems of this type.

4.3.2 Results

Task Success

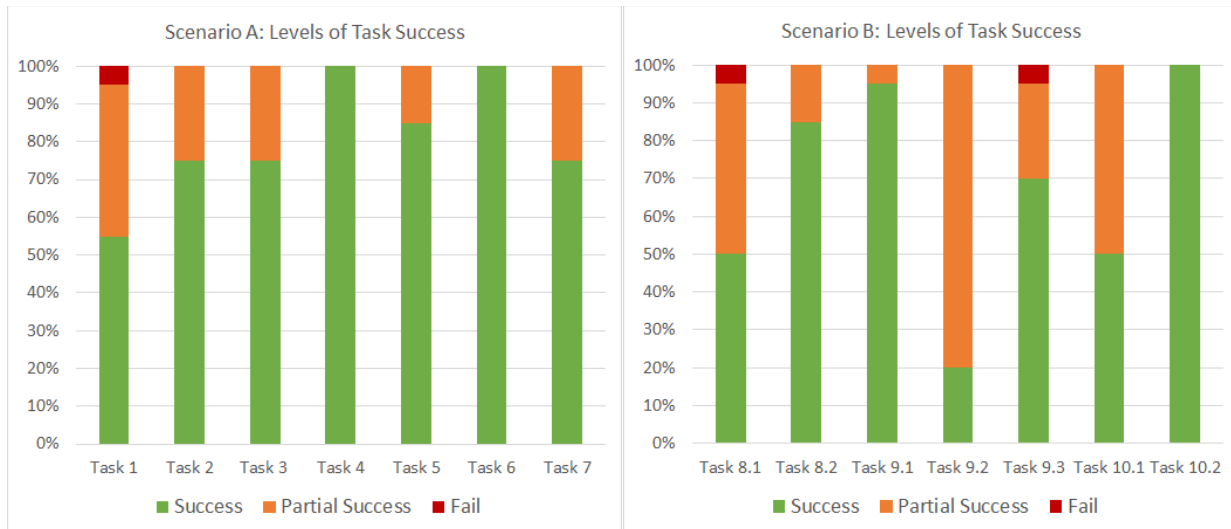


Figure 40 Task Success for scenario A (left) and B (right)

Figure 40 illustrates the task success of scenarios A and B. Each chart is made up of an x-axis that depicts each task in the scenario and a y-axis that represents the users' success in completing the task. There are three types of outcomes for each task: "success" where the users completed the task on their own, "partial success" where the users completed the task but either requested help from the evaluator or were stuck for a significant amount of time and the evaluator decided that they should be offered help. Finally, there is the "fail" outcome, which means that the evaluator had to point out the exact steps to the users in order for them to complete the task.

As shown in Figure 40, almost all of the tasks in both scenarios, with the exception of task 9.2, had a "success" percentage of 50% or higher. Tasks 4, 6, and 10.2 had 100 percent of users complete them successfully on their own, while nine tasks from both scenarios had a "success" outcome of more than 70%. In general, scenario A, which was composed of tasks related to dashboard comprehension, had a higher "success" outcome than scenario B, which was primarily focused on dataset combining and dashboard creation. There are three tasks in particular with "success" outcomes of 50% or lower. These tasks belonged to scenario B and had to do with the creation of a dataset combination, the configuration of a dashboard's data drilling visualization,

and the configuration of a dashboard's views. Additional details on the issues participants faced while trying to accomplish these tasks, are provided in the sections that follow.

Help Requests

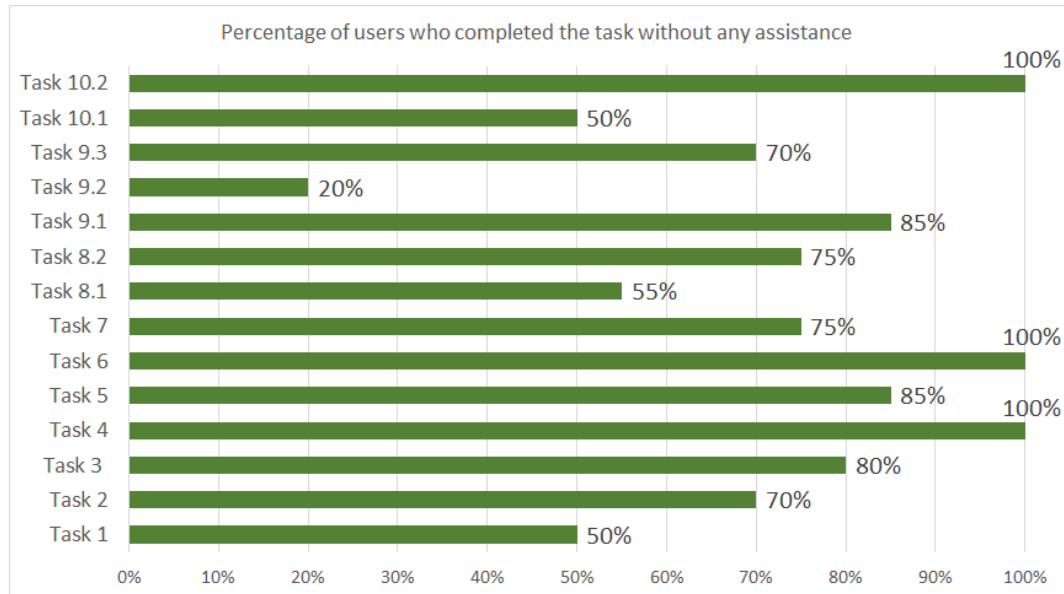


Figure 41 Percentage of users who completed the task without any assistance

Figure 41 shows the percentage of users who completed the task without the assistance of the evaluator. The x-axis shows the percentage of users, while the y-axis shows all of the tasks from both scenarios. In Figure 41 it is illustrated that, with the exception of task 9.2, the percentage of users completing both scenarios' tasks without any assistance was 50% or higher, a percentage corresponding to the task success of those tasks on Figure 40. Similarly, the tasks completed by 70% or more of the users, as well as the tasks completed by all users without any assistance, had nearly the same percentage as the task success of those tasks in Figure 40. Lastly for task 9.2 we can observe that only 20% of the users manage to complete it without the assistance of the evaluator, a fact that can also be noticed on the success and partial success percentages of that task in Figure 40.

Errors observed

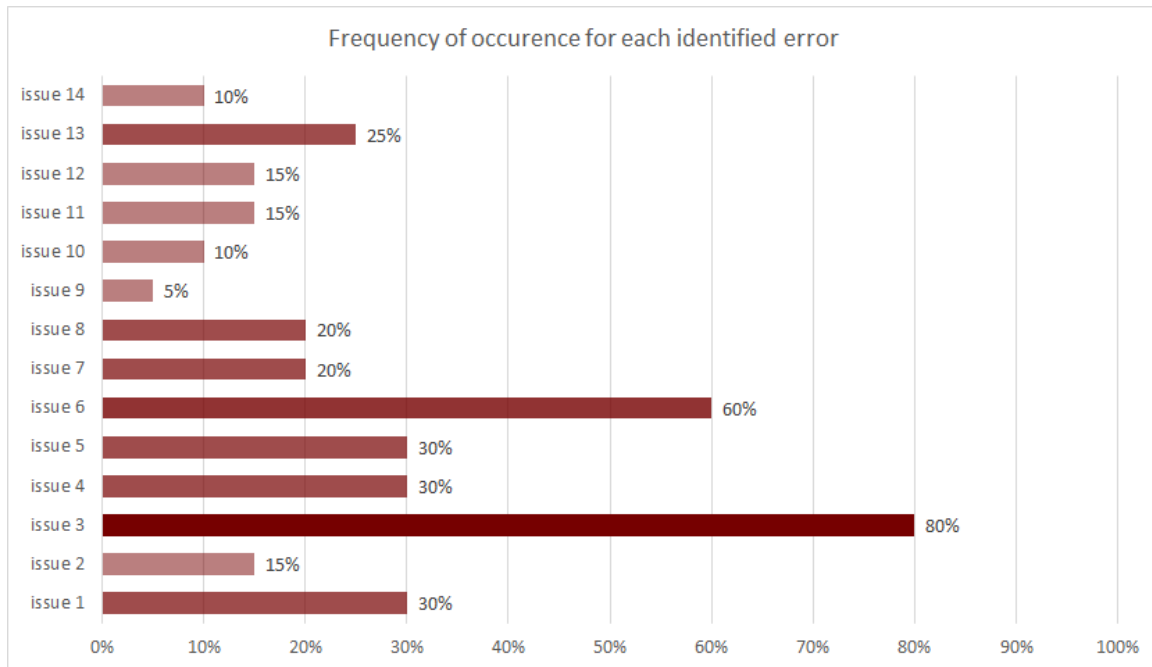


Figure 42 Frequency of occurrence for each identified error during users' scenario executions

Figure 42 shows all of the user errors as well as the frequency of each error. The x-axis shows the percentage of users who made the error, and the y-axis shows the issue number. In summary, the most frequent errors were about some aspects of the dashboard and view creation as well as in the datasets' combining process. A table follows that describes each of the observed issues as well as the scenario task(s) in which those errors were observed.

Table 4 Discovered Issues Description

| Issue | Description | Task(s) |
|---------|---|-------------|
| Issue 1 | When a selected area on the map becomes unselected (by moving the map) users don't notice. | Tasks 3 & 5 |
| Issue 2 | No-data areas on the map are the same color as the lowest range. | Task 4 |
| Issue 3 | While creating a dashboard and configuring the map, users don't understand what the Geospatial Attribute is used for. | Task 9.2 |

CHAPTER 4. EVALUATION

| | | |
|----------|--|-----------|
| Issue 4 | Users can't find where the already combined datasets are. | Task 1 |
| Issue 5 | Setting ranges for the map during the configuration. | Task 9.3 |
| Issue 6 | Users don't fill the title field while creating a view, thus not being able to complete the creation. | Task 10.1 |
| Issue 7 | Users can't find a way to change the title of a dataset. | Task 8.2 |
| Issue 8 | The map can be zoomed. | Task 2 |
| Issue 9 | The areas of the map can be clicked. | Task 2 |
| Issue 10 | When trying to select an existing dashboard of a combined dataset, some users selected the "create dashboard" icon on the card. | Task 1 |
| Issue 11 | Some users try to create a combination and forget to choose the attributes for combination. | Task 8.1 |
| Issue 12 | Users can't find the "cancel" button when they want to unselect an incorrect dataset for combination during the combination creation phase. | Task 8.1 |
| Issue 13 | Users are at the region level on the map and try to see statistics of an area that is at a municipality level (trying to select a municipality from region map level). | Task 7 |
| Issue 14 | Users are not sure if the icon selection during the dashboard creation is of any specific importance. | Task 8.1 |

Overall, it turns out that the most common sources of error pertain to forms, which have to be filled-in with specialized information. Being first time users, participants were not able to fill-in such forms without assistance.

Workload

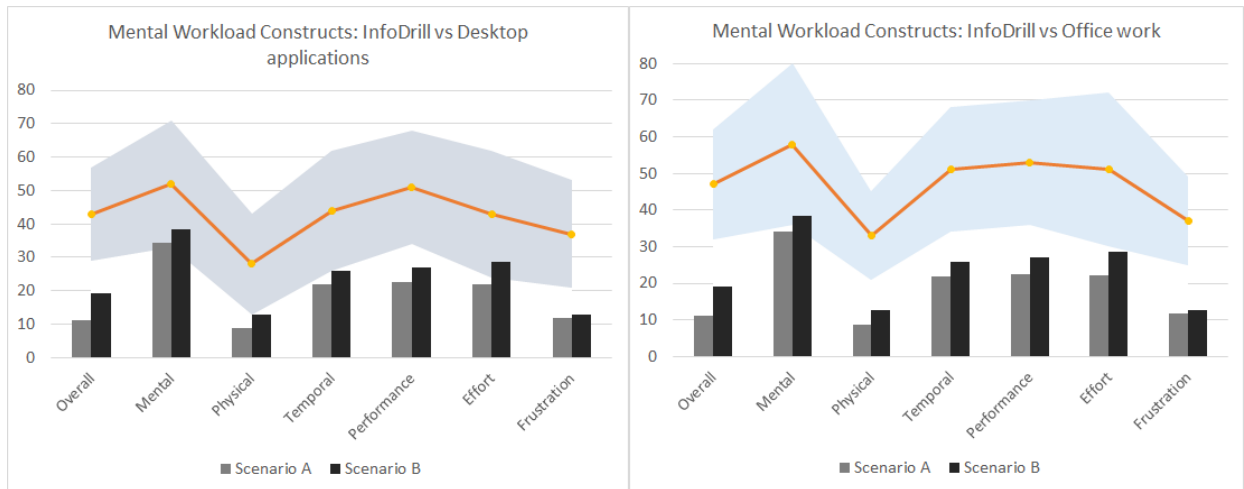


Figure 43 Mental Workload of InfoDrill Application for scenarios A & B versus Desktop applications (left) and Office work in general (right) [48]

Figure 43 shows the mental workload of the InfoDrill application and the impact it had on the users in both scenarios A and B. As we can see, scenario A, which dealt with dashboard comprehension, had a lower average workload and scored lower on all six subscales than scenario B, which dealt with dataset combination and dashboard creation.

The workloads of scenarios A and B were also compared to workloads of desktop applications and office work in general, which were sourced from Hertzum's [48] meta-analytic review of 556 studies on workloads measured with the TLX and its six subscales. In this review, the studies were divided into groups, and the mean of the TLX value and its subscales, as well as the standard deviation of these values, were calculated for each group. For our evaluation, we compared the mean and standard deviation of the desktop applications and office work groups to our results.

As shown in Figure 43, the InfoDrill application is below the mean (indicated by the red line) of both desktop applications and office work, both in terms of overall workload and for all TLX subscales. Some of the subscales, such as mental demands and effort, are approaching the mean of both workloads, but remain within the standard deviation. As a result, it can be concluded that InfoDrill does not induce extraneous workload to users, not even in the case of Scenario B, featuring tasks which were more difficult to be accomplished.

User Experience

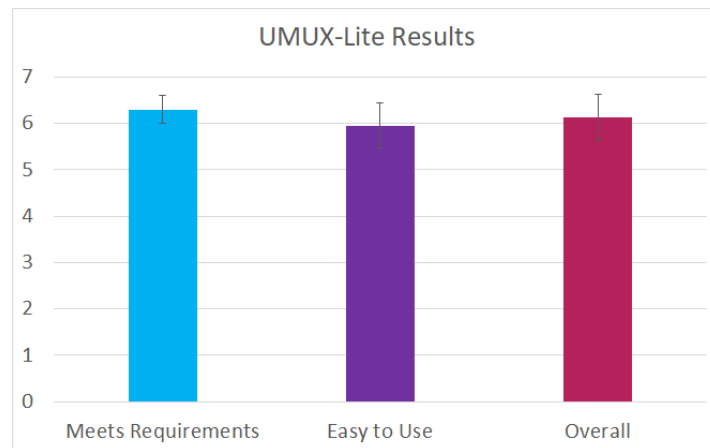


Figure 44 UMUX-Lite Results (with error bars representing 95% confidence intervals)

Figure 44 shows the results of the UMUX-Lite questionnaire, which was completed by users at the end of their session and focused on their User Experience. The results for both the "Meets Requirements" and "Easy to Use" constructs were highly positive, with average values close to and slightly higher than six (6) and with an overall UMUX-Lite result of 6.12 which, when translated with the use of a translation table from [51] it leads to a SUS score of 78.42. This score is generally considered as a good score since it is higher than the 68-point SUS benchmark that was set by the average SUS score collected from 500 studies by Sauro [52].

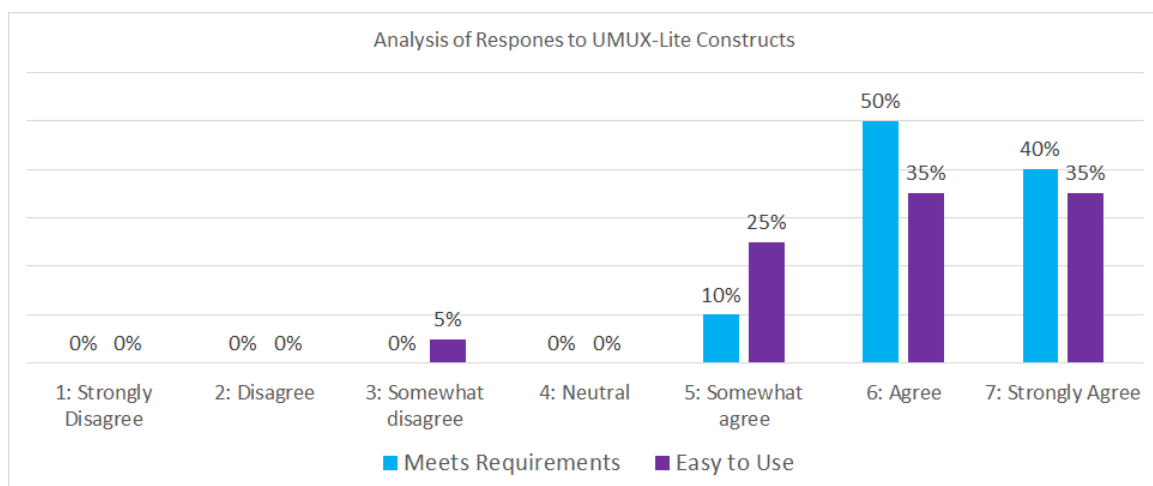


Figure 45 UMUX-Lite Analysis of Responses

Figure 45 shows the analysis of UMUX-Lite questionnaire responses. Almost all of the responses for both constructs were positive (either 5, 6, or 7, denoting "Somewhat agree," "Agree," or "Strongly Agree" respectively). The most common response for the "Meets Requirements" construct was 6, with 50% of users giving this response, followed by answers of 7 and 5. The most common responses for the "Easy to Use" construct were 6 and 7, with 35% each, followed by 5 and, finally, 3 ("Somewhat disagree"), which was a response given by 5% of users. As a result, it can be concluded that the overall UX was positive for almost all users. In particular, it was unanimously agreed that the system met users' requirements, while the majority of participants also agreed that it was easy to use.

Debriefing findings

At the end of each session, users were asked three brief questions with regards to their experience and familiarity with similar information visualization systems. For the first question regarding what the users liked the most about the application, most of the answers revolved around the ease of use of the application as well as its importance at assisting in decision-making.

For the question on what should be changed or improved in the application, about 35% of the users didn't consider that something should be changed. The rest gave the following suggestions:

- Need to add more advanced comparisons and customizing of parameters (20% of users)
- Need for explanation text on each step of the forms for creating a dashboard or view (10% of users)
- Need for more language options (specifically Greek) (10% of users)
- When configuring a visualization, provide more than just the information about the data type of an attribute (5% of users)
- The map needs a different color palette so that the areas in different categories can be distinguished more easily. (5% of users)
- Add ability to create user-defined dataset categories (5% of users)
- More types of data analysis (5% of users)
- Ability to import data sources directly from other databases (5% of users)
- Ability to share a dashboard (5% of users)

And for the last question on whether the users had any prior experience with a similar application the answers were as follows:

- No prior experience with a similar application, but have statistics experience. (55% of users)
- Little prior experience with a similar application, occasional use of Excel spreadsheets for creation of charts and graphs (5% of users)
- Some experience with a similar application, use of custom data visualization solutions for specific use cases (20% of users)
- Experience with a similar application, use of systems like PowerBI (20% of users)

4.4 Discussion

For the framework's evaluation study two evaluations were conducted. A heuristic evaluation of the DaRAV application and a user testing evaluation for the InfoDrill application. Both of these evaluations brought forward some interesting results. In terms of the DaRAV evaluation, the reporting of three UX experts revealed 23 issues. Those issues ranged in severity from cosmetic to major (based on the scale mentioned in section 4.1.1), with 18 of them being of cosmetic or minor severity and 5 of them being of major severity. These major severity application problems were primarily found on the forms that a user must fill out in order to create a new risk analysis or anonymization process, and specifically had to do with field labeling, input methods, and error messages. Although these issues are not considered as show stoppers, they are already being addressed in the next version of the application. Beside those issues though, the overall impression of the experts with regards to the user experience of the application was positive.

The user testing evaluation of the InfoDrill application was carried out with users from various domains and with varying levels of expertise. The testing consisted of two scenarios: one for dashboard comprehension and the other for dashboard creation, as well as dataset combining. The evaluation results revealed that the success and partial success percentages of scenarios' tasks completed by users were high for both scenarios of the evaluation. The scenario where the users did the least amount of errors and subsequently needed the lesser amount of help was the first one. In terms of the workload experienced by users during the execution of both scenarios, the evaluation results revealed that the overall workload was significantly lower, even when compared to other desktop applications or office work workload studies. The application's user experience (UX), as perceived by users, was considered to be very good, with the average results of the UMUX-Lite questionnaire being 6.12, equivalent to a SUS score of 78.42. This is considered a very good score since it is higher than the 68-point SUS benchmark for user experience.

Chapter 5

Conclusions and Future Work

In this chapter, we provide a summary of the thesis' work as well as our scientific contributions. Additionally, we discuss about future work and the ways in which the proposed framework could be enhanced.

5.1 Conclusions

This thesis provided a unified framework aimed at the analysis, visualization, and exploration of big data while ensuring security and privacy. The proposed framework is realized by two applications. The first application is DaRAV, which provides a platform that can facilitate de-anonymization risk analysis modules and anonymization functionalities. This application advances the state of the art by providing a wide range of de-anonymization risk analysis options for datasets that contain more than just tabular data. DaRAV, in particular, provides risk analysis modules for tabular, spatiotemporal, textual, financial transactions, and aggregation-based data, thus assisting data owners in visualizing and analyzing the risk of leaking personal data that may pass through their dataset. DaRAV also provides anonymization capabilities, allowing users to perform k -anonymity and l -diversity anonymization processes on their datasets.

The second application of the framework is InfoDrill, which provides a solution for visualizing and exploring large datasets by combining previously owned datasets with those obtained from digital data marketplaces and displaying them through smart interactive dashboards. Our scientific contribution to the industry standard visualization of data through dashboards is that this application allows users to create dashboards and populate them with visualizations that can facilitate data drill down and roll up actions on a combination of two datasets. These actions can

be applied to all of the dashboard's visualizations without any additional configuration. This approach, reduces the visual noise that users may encounter when dealing with large datasets since it presents data in hierarchies that users can search through using drill down or roll up actions without being in danger of losing the large image perception of the data under analysis. In this way, the application gives to the users the ability to explore through data, discover new knowledge, and gain new insights that they could not have gained without performing a similar analysis on those datasets.

Both of the framework's applications were evaluated for their User Experience (UX) and their ease of use and workload (in the case of InfoDrill) in the framework's evaluation study. Specifically, a heuristic evaluation was conducted for the DaRAV application from three UX experts and a user testing evaluation for the InfoDrill application with 20 executive, marketing, data analysis and technical personnel from telecommunications and financial companies as well as local municipalities. The heuristic evaluation of DaRAV resulted in an evaluation report containing User Experience (UX) issues that should be addressed, the majority of which were minor or cosmetic in nature. InfoDrill's user testing resulted in a very good User Experience, with a SUS score of 78.42, which is higher than the 68-point SUS benchmark. Furthermore, when compared to other desktop applications or office work workload studies, the users' perceived workload while using the InfoDrill application was significantly lower. These evaluation results indicate that users can effectively and efficiently combine datasets and create dashboards, but also successfully utilize dashboards to comprehend information, drilling-down and rolling up to different levels of detail, according to their current analysis needs, by merely handling the "master" dashboard visualization (e.g. the geospatial data visualization).

5.2 Future work

The positive findings of the framework's evaluation study have strengthened our motivation for the further development of this framework. One of the first priorities with regards to future work would be the addressing of all the issues discovered in both the heuristic and users testing evaluations of the applications. Afterwards, a user testing evaluation for the DaRAV application would be highly beneficial.

CHAPTER 5. CONCLUSIONS AND FUTURE WORK

For the further advancement of the framework, with regards to the DaRAV application, the addition of more advanced data anonymization functionalities as well as the export of interactive risk analysis reports could be of great assistance to the users. Furthermore, the addition of data utility analysis to the application would be a feature that could greatly benefit data owners because it would assist them in determining the level of anonymity they would like to achieve by taking into account the quality and utility of the output data.

Regarding the InfoDrill application, the ability to share already created dashboards in the form of a report as well as the addition of collaborative data analysis would be highly beneficial. Additionally, a future endeavor would be to augment the application's current descriptive analysis with the addition of predictive and prescriptive ones. This could be accomplished by incorporating predictive modeling and machine learning, which would analyze current and historical data in order to make a better assessment of what will happen in the future and provide insights that could greatly benefit the decision-making process.

References

- [1] EU, "General Data Protection Regulation," [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
- [2] A. Bampoulidis, A. Bruni, I. Markopoulos and M. Lupu, "Practice and Challenges of (De-) Anonymisation for Data Sharing," in *International Conference on Research Challenges in Information Science*, 2020.
- [3] L. Fu and W.-C. Hu, *Online Analytical Processing and Data-Cube Technologies*, IGI Global, 2008.
- [4] L. Sweeney, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.
- [5] L. Sweeney, "Uniqueness of Simple Demographics in the U.S. Population," Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.
- [6] S. Hansell, "AOL Removes Search Data on Group of Web Users," *NYTimes*, 2006. [Online]. Available: <http://www.nytimes.com/2006/08/08/business/media/08aol.html>.
- [7] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *IEEE Symposium on Security and Privacy*, 2008.
- [8] D.-F. Josep, S. David and S.-C. Jordi, *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*, Morgan & Claypool, 2016.
- [9] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," 1998.
- [10] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkatasubramanian, "l-diversity: Privacy beyond k-anonymity.," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, p. 3–es, 2007.
- [11] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007.
- [12] C. Dwork, "Differential Privacy," in *International Colloquium on Automata, Languages, and Programming*, 2006.
- [13] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, C. B. Pierce and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *IEEE 27th Computer Security Foundations Symposium*, 2014.

REFERENCES

- [14] L. Tiancheng and L. Ninghui, "On the tradeoff between privacy and utility in data publishing," in *15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.
- [15] F. Prasser, J. Eicher, H. Spengler, R. Bild and A. K. Kuhn, "Flexible data anonymization using ARX - Current status and challenges ahead," *Software: Practice and Experience*, 2020.
- [16] "ARX Data Anonymization Tool," [Online]. Available: <https://arx.deidentifier.org/>.
- [17] T. Rolandus Hagedoorn, R. Kumar and F. Bonchi, "X2R2: a tool for explainable and explorative reidentification risk analysis," *Proc. VLDB Endow.*, 2020.
- [18] D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann, *Mastering The Information Age – Solving Problems with Visual Analytics*, 2010.
- [19] G. Margetis, S. Ntoa, M. Antona and C. Stephanidis, "HUMAN-CENTERED DESIGN OF ARTIFICIAL INTELLIGENCE," in *Handbook of Human Factors and Ergonomics, Fifth Edition*, 2021, pp. 1085-1106.
- [20] A. M. Peterson and R. Kimchi, *Perceptual Organization in Vision*, Oxford Handbooks Online, 2013.
- [21] A. Treisman, "Preattentive processing in vision.," *Comput. Vision Graph. Image Process.*, p. 156–177, 1985.
- [22] J. J. Thomas and K. A. Cook, "Illuminating the Path: The Research and Development Agenda for Visual Analytics," *IEEE CS Press*, 2005.
- [23] A. B. M. Moniruzzaman and A. S. Hossain, "NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison.," *Int J Database Theor Appl.* 6, 2013.
- [24] A. E. Brewer, "Towards robust distributed systems (Invited Talk)," *Principles of Distributed Computing*, Portland, Oregon, 2000.
- [25] M. Khan and S. Khan, "Data and Information Visualization Methods, and Interactive Mechanisms: A Survey.," in *International Journal of Computer Applications*, 2011.
- [26] E. Gorodov and V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data," *Journal of Electrical and Computer Engineering*, 2013.
- [27] S. M. Hajirahimova and I. M. Ismayilova, "BIG DATA VISUALIZATION: EXISTING APPROACHES AND PROBLEMS," *Problems of Information Technology*, 2018.
- [28] J. Moura and C. Serrao, "Security and Privacy Issues of Big Data," 2015.
- [29] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "Knowledge discovery and data mining: towards a unifying framework.," in *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, 1996.

REFERENCES

- [30] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis and D. A. Keim, "Knowledge Generation Model for Visual Analytics," in *IEEE Transactions on Visualization and Computer Graphics*, 2014.
- [31] "Tableau Desktop," [Online]. Available: <https://www.tableau.com/>.
- [32] "Elasticsearch," [Online]. Available: <https://www.elastic.co/what-is/elasticsearch>.
- [33] "Power BI," [Online]. Available: <https://powerbi.microsoft.com/en-us/>.
- [34] "TRUSTS - Trusted Secure Data Sharing Space," [Online]. Available: <https://www.trusts-data.eu/>.
- [35] "Safe-DEED - Safe Data-Enabled Economic Development," [Online]. Available: <https://safe-deed.eu/>.
- [36] "Angular," [Online]. Available: <https://angular.io/>.
- [37] "Leaflet," [Online]. Available: <https://leafletjs.com/>.
- [38] "plotly," [Online]. Available: <https://plotly.com/javascript/>.
- [39] "Spring Boot," [Online]. Available: <https://spring.io/projects/spring-boot>.
- [40] "Spring Boot Maven Plugin Documentation," [Online]. Available: <https://docs.spring.io/spring-boot/docs/current/maven-plugin/reference/htmlsingle/>. [Accessed 17 12 2021].
- [41] "SQLite," [Online]. Available: <https://www.sqlite.org/index.html>.
- [42] "Docker," [Online]. Available: <https://www.docker.com/>.
- [43] "Logstash," [Online]. Available: <https://elastic.co/logstash>.
- [44] "D3js," [Online]. Available: <https://d3js.org/>.
- [45] "Google Charts," [Online]. Available: <https://developers.google.com/chart>.
- [46] "Node.js," [Online]. Available: <https://nodejs.org/en/>.
- [47] J. Nielsen, *Usability Engineering*, Elsevier, 1994.
- [48] M. Hertzum, "Reference values and subscale patterns for the task load index (TLX): a meta-analytic review," *Ergonomics, Taylor & Francis Online*, vol. 64, no. 7, pp. 869 - 878, 2021.
- [49] K. Finstad, "The Usability Metric for User Experience," *Interacting with Computers*, vol. 22, pp. 323-327, 2010.
- [50] J. Lewis, B. Utesch and D. Maher, "UMUX-LITE: when there's no time for the SUS," in *Conference on Human Factors in Computing Systems - Proceedings*, 2013.
- [51] J. R. Lewis, "The System Usability Scale: Past, Present, and Future," *International Journal of Human-Computer Interaction*, vol. 34, no. 7, pp. 557-590, 2018.

REFERENCES

- [52] J. Sauro, "Measuring Usability with the System Usability Scale (SUS)," [Online]. Available: <https://measuringu.com/sus/>. [Accessed 16 12 2021].