Leveraging users' behavior, intentions and interests for enhancing Exploratory Data Analysis and Data Narration

Veronika Peralta

HABILITATION À DIRIGER DES RECHERCHES

Discipline : Informatique

Année universitaire : 2023 / 2024

présentée et soutenue publiquement par :

Verónika Peralta

5 avril 2024

Leveraging users' behavior, intentions and interests for enhancing Exploratory Data Analysis and Data Narration

JURY :

Mme. Sihem AMER YAHIA	Directrice de Recherche, CNRS	Université de Grenoble Alpes
Mme. Karine BENNIS ZETOUNI	Professeure des universités	Université Paris-Saclay
M. Jérôme DARMONT	Professeur des universités	Université de Lyon 2
M. Thomas DEVOGELE	Professeur des universités	Université de Tours
Mme. Rokia MISSAOUI	Professeure	Université du Québec en Outaouais
M. Arnaud SOULET	Professeur des universités	Université de Tours

Acknowledgements

First of all, I express all my gratitude to Professors Sihem Amer Yahia, Karine Benis Zetouni and Rokia Missaoui who did me the honor of reviewing my dissertation. I thank them for taking the time to report my work despite their many obligations. Their very relevant and constructive reports and comments led to very interesting discussions during the defense and opened up new perspectives. I am also very honored by the presence in the jury of Professors Jérôme Darmont, Arnaud Soulet, Patrick Marcel and Thomas Devogele. I thank them for their careful reading of this work and for the valuable feedback they gave me. I acknowledge Jérôme Darmont for agreeing to be president of the jury and for brightening up the defense in hybrid mode, in addition to the quality of his remarks and questions. I am deeply grateful to Thomas Devogele and Patrick Marcel for their generous and unwavering support. Beyond their scientific qualities and open-mindedness, they offered me the opportunity to carry out a rich and varied research activity.

The contributions of this dissertation are above all the synthesis of a team work. I particularly thank all the PhD students - old and current - that I had the privilege of advising: Mahfoud Djedaini, Krista Drushku, Frederick Bisone, Clément Moreau, Raymond Ondzigue Mbenga, Faten El Outa, Flavia Serra, Raphaël Bres, Guillaume Tejedor y Dominguez, Hiba Merakchi and John Dawson. I thank Patrick Marcel, Nicolas Labroche, Thomas Devogele, Laurent Etienne, Edgar Ngoungou, Adriana Marotta, Cyril de Runz, Anna Maria Raimond and Arnaud Le Guilcher, for trusting me to supervise these theses. I learned a lot alongside them and had a lot of fun working with them. I know that other collaborations await us. I acknowledge Hélène Blasco for agreeing to supervise Guillaume's thesis. I also thank Kamal Boulil, Louise Parkin and Alexandre Chanson for all the fruitful exchanges we had during their postdoc and postgraduate projects, as well as the multitude of Master and Bachelor students who directly or indirectly contributed to this work.

I had the chance to collaborate directly with many colleagues who belong or belonged to the BDTLN team. I thank them all without forgetting the technical and administrative teams, and the other colleagues from the CS department and the IUT who make Blois a friendly and fulfilling working environment.

This dissertation also owes a lot to the numerous exchanges resulting from regional to international collaborations, short or long-term, but always enriching. I would particularly like to thank Panos Vassiliadis, Dimos Gkitsakis, Stefano Rizzi and Matteo Francia.

I want to thank Mokrane Bouzeghoub and Raúl Ruggia, my PhD thesis supervisors, for inspiring me and transmitting their passion for research, rigor and perseverance. I thank them for the trust they have placed in me. Thanks also to Zoubida Kedad, with whom I had the opportunity to work during and after my thesis and I hope that our collaborations continue.

Beyond the scope of work, I would like to thank my family and friends from Uruguay, who attended my entire defense, despite the language barrier. I would like to extend my thanks to David, Emilie and Gabriel who accompany me in life. I wrote this dissertation over many weekends and vacations and I thank them for their understanding and unwavering support.

Abstract

Exploratory Data Analysis (EDA) is an analysis technique used for efficiently extracting knowledge from data even when we do not know exactly what we are looking for. EDA is at the core of Data Narration (DN), the process of narrating data stories supported by data analysis. While much research effort is put in the automation of EDA and DN, users' behavior, intentions and interests are frequently neglected, leading to fixed not-personalized data reporting and storytelling. This Habilitation thesis is a contribution to the huge task of developing user-centric EDA and therefore intentional DN.

We firstly propose techniques for learning users' analysis behavior from query workloads. We segment large query workloads into explorations, i.e. coherent sequences of queries related to a same information need. We propose classification models to evaluate to what extent a query is focused and contributes to the success of an exploration, a Knowledge Tracing model to assess users' analysis skills, and a clustering method to group explorations revealing similar analysis patterns, i.e. sharing similar sequences of operations and containing queries of close complexity. Our methods rely on a model of queries and explorations from the prism of users' skills, based on a large set of features capturing various aspects of a query and its context within the exploration, in particular, query fragments, operations and timing. A similarity measure tailored for explorations allows the discovery of analysis patterns translating users' behavior.

We then turn to user's interests. We propose a two-level framework for developing interestingness measures, consisting respectively of high-level interestingness aspects, and data-oriented assessment algorithms. Focusing in a particular interestingness aspect (the relevance of a query for the overall analysis intention of the user), we propose an approach for learning user interests in a query workload and recommending relevant queries. We formalize the problem of discovering coherent user interests as a clustering problem, for which a similarity measure is learned to capture whether two queries reflect a same user interest. To leverage the discovered user interests for the purpose of query recommendation, we propose an original interest-based recommender.

We eventually consider EDA within the DN process. As apart some general considerations, there is no consensual definition of DN, let alone a model of it, we start by proposing a conceptual model that provides a structured, principled definition of the key concepts of the domain. We then incorporate dynamic aspects and propose a process model that covers the whole DN cycle and accommodates a wide range of practices observed in the field. Both models draw attention to the importance of EDA support and highlight intentional aspects.

Finally, this dissertation discusses several research perspectives. This work is undertaken within the framework of 10 PhD theses and 7 research projects.

Keywords: Exploratory Data Analysis, Data Narration, User behavior, User interests, User intentions

Contents

Chapter 1

Introduction

This dissertation presents a synthesis of the research that I carried out at the LIFAT Laboratory of the University of Tours in the area of Exploratory Data Analysis over the period 2014 – 2023. In this dissertation, I focus on three aspects that I treated aiming to better support data exploration: learning users' analysis behavior, leveraging users' interests, and modeling data narratives.

This first chapter describes the context, challenges, scientific objectives and general approach of my research work.

1.1 Context: From *data*, to *insights*, to *data narratives*

The exponential grow of available data is now a well-known phenomenon. 5G connectivity, sensors everywhere, connected devices, and the shift to cloud and edge computing are just a few forces that are seeing the volume of data produced globally increase fivefold between 2018 and 2025, from 33 to 175 zettabytes (10^{21} bytes), according to [European Commission et al., 2020].

Amidst this torrent of data, it is of paramount importance to be able to access and exploit relevant data efficiently. Indeed, people and organizations need to collect, clean, transform, integrate, store, explore, analyze and summarize such amounts of data in order to gain insights and support any data-driven decision-making. An **insight** is the understanding of a particular cause and effect based on the identification of relationships and behaviors within a particular context[1]. It is a deep form of knowledge, an accurate and deep understanding of something. But insight extraction is not the final goal. Insights should be reported and communicated, to make it easy for stakeholders to interpret, understand and act on the data being shared. A **data narrative**, i.e. a story, supported by facts extracted from data analysis, and rendered using interactive visualizations [Carpendale et al., 2016], is the prominent communication media.

To cope with this data deluge, a plethora of solutions arose from diverse overlapping research fields, aiming to integrate heterogeneous data from diverse sources, summarize indicators, discover patterns, draw inferences, make predictions and simulations, gain competitive business advantage, and support strategic decision-making. Without being exhaustive, we can mention:

- Data management: the process of ingesting, storing, organizing and maintaining the data created and collected by an organization [Watson, 2006, Abiteboul et al., 2018],
- Business Intelligence: tools and techniques that transform business data into timely and accurate information for decision-making [Chaudhuri et al., 2011, Rizzi, 2018],

[1] `https://en.wikipedia.org/wiki/Insight`

- Data Mining: the process of discovering knowledge or patterns from massive amounts of data [Han et al., 2011, Gupta and Chandra, 2020],
- Machine Learning: algorithms that allow computer programs to automatically improve through experience [Shalev-Shwartz and Ben-David, 2014, Sarker, 2021],
- Visual Analytics: the science of analytical reasoning supported by interactive visual interfaces [Keim et al., 2018, Yuan et al., 2021],
- Big Data Analytics: technical means dealing with both theory and application of big data [Markl et al., 2018, Shi, 2022], and
- Cognitive Analytics: processes and algorithms that mimic human cognitive processes [Gudivada et al., 2016].

Over the past decade, **Data Science** has emerged as a major interdisciplinary field and its use drives important decisions in enterprises and discoveries in science [Abadi et al., 2022]. Data Science is defined as "the processes and systems that enable the extraction of knowledge or insights from data in various forms, either structured or unstructured"[2]. Data Science process are generally conceived as workflows, combining multiple tasks to gather raw data from multiple sources, analyze it, and present the results in an understandable format. Such chaining of tasks is called a **pipeline**. We remark that many of those tasks are based on techniques developed in the related fields listed above, Data Science having much overlap with them.

De Bie et al. studied Data Science from the prism of automation, and organized Data Science activities in four quadrants, shown in Figure 1.1, namely: Data Engineering, Data Exploration, Model Building and Exploitation [De Bie et al., 2022]. They highlight that important parts of Data Science are already being automated, especially in the model building stages, where techniques such as automated machine learning (AutoML) are gaining traction, but stress that other aspects are harder to automate, not only because of technological challenges, but because open-ended and context-dependent tasks require human interaction.

Another related field is **Data Narration** (DN), where vast effort is put beforehand, in data collection, wrangling and integration, as well as afterwards, in data reporting, visualization and storytelling. DN is much more than exploring data and reporting results, it is a matter of narrating a story supported by data, conveying insights to an intended audience, typically via powerful visualizations [Carpendale et al., 2016]. Actually, DN can be seen as a particular Data Science pipeline, its tasks covering the four quadrants of Figure 1.1.

The data deluge also triggered a commercial success in the Software Industry, and many techniques and tools arose under the general name of **Data Analytics**, which also includes (and frequently confounds) many of the fields listed above. As reported by Gartner, increasingly Data Analytics has become a primary driver of business strategy, and the potential for data-driven business strategies and information products is greater than ever. They forecast that Data Analytics, Business Intelligence, Data Science and Machine Learning continue to collide, driving advanced Business Intelligence and Data Science-Machine Learning platform consolidation. By 2026, 50% of Business Intelligence tools will activate their users' metadata, offering insights and data narratives with recommended contextualized journeys and actions [James and Duncan, 2023].

Undoubtedly, data holds the key to create business value and fuel company success. Nevertheless, even having a lot of technology at fingertips, gaining insights and supporting decision-making is a very complex process. While decision-makers dream of insights, they frequently just

[2]U.S. National Science Foundation, Computer and Information Science and Engineering (CISE) – `https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf`

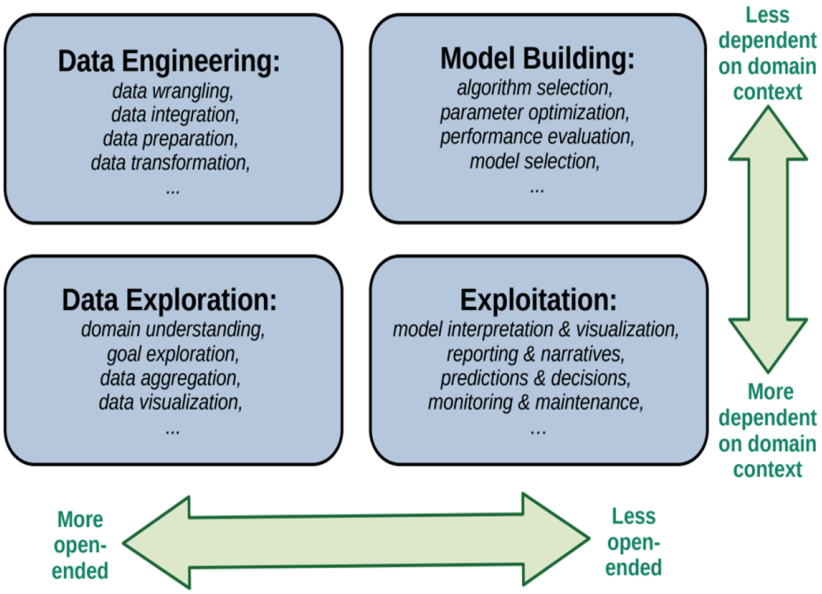


Figure 1.1: Data Science activities organized in 4 quadrants, taken from [De Bie et al., 2022]

obtain some factual knowledge. Chief Information Officers and Senior Data Analysts worldwide testify that "turning data into actionable information is not an easy journey"[3] and "the latest algorithms do not get off the hook for making decisions altogether"[4]. In addition, "Data Analytics can unearth new questions, as well as innovative solutions and opportunities that business leaders had not yet considered"[5]. Unlocking data's full potential relies on a complex process, demanding large data exploration, deep understanding and sound data analysis.

This dissertation places at the crossroad of the mentioned fields. While considering the whole Data Narration process, and thus the four quadrants of Figure 1.1, special attention is paid to Data Exploration.

1.2 Challenges

Exploratory Data Analysis (EDA), Interactive Data Exploration (IDE), or simply Data Exploration, is about efficiently extracting knowledge from data even if we do not know exactly what we are looking for [Idreos et al., 2015]. EDA was introduced by John W. Turkey, who shows how simple graphical and quantitative techniques can be used to open-mindedly explore data. According to Tukey, EDA is about hypothesis generation. Unlike confirmatory analysis, EDA uses data to identify potential hypothesis to explain observed phenomena and assist with selection of appropriate statistical tests [Turkey, 1977].

On-line Analytical Processing (OLAP) is a good example of EDA. OLAP is a major Business Intelligence technique, providing navigation through data to non-expert users, so that they are able to interactively generate ad hoc queries without the intervention of IT professionals [Abelló and Romero, 2018]. OLAP data (the facts to be analyzed, e.g. sales) are described by a set of dimensions (the analyses axes, e.g. product, customer, time) and a set of measures (the numeric indicators, e.g. sale amount). They are structured in a multidimensional space produced by the set of dimensions, called a datacube (or simply a cube). OLAP takes advantage of simple

[3]Deloitte – `https://www2.deloitte.com/cy/en/pages/technology/articles/data-grown-big-value.html`
[4]Forrester – `https://www.forrester.com/what-it-means/ep328-techs-role-in-decisions/`
[5]Gartner – `https://www.gartner.com/en/information-technology/insights/data-analytics`

primitives like drill-down or slice-and-dice for the navigation in the multidimensional space. For example, an analyst may explore several measures and cross several dimensions, in order to find clues, causes or correlations to explain unexpected data values, until identifying the most relevant data subsets and deeply analyzing them.

Nowadays, EDA spreads beyond Data Management and Business Intelligence frontiers, and places at the core of many processes and systems that enable the extraction of knowledge or insights from data in various forms, in particular, Data Science and Machine Learning pipelines. Specifically, Data Exploration and Model Building tasks are frequently combined, pipelined or merged, thus the frontiers of the corresponding quadrants of Figure 1.1 are disappearing. In this sense, the Seatle Report on Database Research claims that "Data Science and Database Research communities must work together closely to enable data to insights pipeline" [Abadi et al., 2022].

Such a joint work is challenging by itself, but, when considering specific pipelines (as DN) or specific application domains (e.g. health or environment), additional challenges appear in terms of **interoperability** and **cross-disciplinarity**. In particular, conceptual models are needed for communication and co-construction with users, as currently, solutions are fragmentary.

EDA support addresses the development of techniques that allow users to explore their data and help them to better gain insights. Such techniques allow the data analyst to look at data to see what it seems to say, uncover underlying structures, isolate important variables, detect outliers and other anomalies, and suggest suitable models for conventional statistics [Hinterberger, 2018].

Many Business Intelligence and Visual Analytics tools propose advanced query interfaces for explore data. But nice query interfaces are not enough if users cannot easily gain insights on the analysed data. A study interviewing 18 data analysts [Wongsuphasawat et al., 2019] found that "Analysts must perform repetitive tasks (e.g., examine numerous variables), yet they may have limited time or lack domain knowledge to explore data. Analysts also often have to consult other stakeholders and oscillate between exploration and other tasks, such as acquiring and wrangling additional data."

To tackle these issues, many works propose techniques for aiding in query formulation and result interpretation. Without trying to be exhaustive, we mention proposals to personalize and recommend queries (see e.g. [Milo and Somech, 2018, Meduri et al., 2021, Lai et al., 2023, Francia et al., 2023]), compose queries to guide the exploration (see e.g. [Bar El et al., 2020, Zolaktaf et al., 2020, Personnaz et al., 2021]) and more generally mining data for completing and highlighting query answers (see e.g. [Aufaure et al., 2013a, Vassiliadis et al., 2019]). We point out recent contributions for collaborative [Sakka et al., 2021b, Muhammad and Darmont, 2023] and conversational [Francia et al., 2022a, Wang et al., 2022] Business Intelligence.

Despite the numerous works for EDA support, there are still many challenges, in particular for developing new exploration techniques for **complex data** (e.g. high-dimensional, sparse, sequential, graph-oriented data) adapting to specific **use-cases** (as done for galaxies [Youngmann et al., 2022]).

Beyond EDA support, **EDA automation** is drawing attention nowadays [Milo and Somech, 2020]. Many recent works address the automatic discovery of insights [Ding et al., 2019, Ma et al., 2021, Ma et al., 2023] and their usage for enhancing data exploration [Bar El et al., 2020, Chanson et al., 2022a] and automating the overall data narration process (e.g. [Wang et al., 2020, Shi et al., 2021, Sun et al., 2023]). Even if such works envision full automation, they only deal with very specific scenarios where data exploration is reduced to the search of statistical findings. Furthermore, **data quality** issues are not addressed, despite being the harder obstacles for data analysts. We think that there is still a long journey to automation.

But, is automation the ultimate goal? We remark, from Figure 1.1, that Data Exploration is the quadrant being the more challenging for automation, beacuse, at pointeb by [De Bie et al., 2022], it requires human interaction. However, the race for EDA automation (and DN automation) is leaving the user behind. A very recent survey on data narration asks for considering cognitive, emotional and contextual impacts [Schröder et al., 2023]. We claim that users' intentions, interests and emotions are frequently neglected, leading to fixed, not personalized, data reporting and storytelling. There is a big need for putting the user in loop, and considering many **intentional** and **contextual** aspects.

1.3 My research experience

My research activities address the general challenge of EDA in the context of complex information systems that integrate data from multiple sources. I am particularly interested in the development of techniques to support EDA, by learning analysts' behavior and interests, and qualifying their analyses, in order to offer tools adapted to their context. I consider the whole data lifecycle, from data collection and quality management, to the restitution of analysis results, narrating a data story.

I can thus classify my research activities into 4 main themes: (i) learning of analyst's behavior and interests, to enhance EDA support, (ii) exploratory analysis of complex data, in particular semantic sequences, (iii) data narration, promoting EDA and analysts' intentions at the core of data narration models, and (iv) data quality management, guided by analyst's needs and context.

Next paragraphs briefly describe my contributions in each theme, undertook in collaboration with several students and colleagues.

Learning analysts' behavior and interests. In order to improve EDA tools, in particular by adapting to analysts' habits, skills and context, we investigated machine learning techniques for learning analysts' behavior, methods and interests, and qualify their analyses.

The PhD thesis of Mahfoud Djedaini [Djedaini, 2017] proposes a benchmark [Djedaini et al., 2016] and learning methods for evaluating the quality of an exploratory analysis [Djedaini et al., 2019]. The PhD thesis of Krista Drushku [Drushku, 2019] focuses on discovering analysts' interests [Drushku et al., 2017] and recommending queries and content based on these interests [Drushku et al., 2019]. In both theses, the analysis of the sequences of queries evaluated by the analysts was capital.

Recent contributions focus on the discovery of exploration patterns and analysts' skills [Moreau et al., 2022], the assessment of the analyst's interests [Gkitsakis et al., 2024], and the balance of performance and interest constraints [Chanson et al., 2020].

Analysis of complex data. The analysis of sequences of complex data (as the sequences of queries analysed in previous theme) revealed itself to be challenging. I am particularly interested in the analysis of semantic sequences (semantic data series). They are sequences of chronologically ordered semantic data, representing various processes (e.g. life courses, daily trips, patient records and flows of varied activities). Their analysis allows the answering of various societal, industrial or individual issues, for example, the detection of dangerous behavior, the detection of difficulties and bottlenecks, and the learning of behavior patterns.

The PhD thesis of Frederick Bisone [Bisone, 2021] studies the trips of connected ambulances and the PhD thesis of Clément Moreau [Moreau, 2021] deals with daily trips of schoolchildren and itineraries of tourists. We proposed semantic enhancement of mobility sequences using multiple sensors [Bisone, 2021], several similarity measures for comparing sequences, which take into

account the semantics of the activities [Moreau et al., 2020c, Moreau et al., 2021b], clustering methods for sequences based on these distances [Moreau et al., 2021a] and cluster analysis techniques [Moreau et al., 2020a].

The starting PhD thesis of Hiba Merakchi aims to extend these proposals, particularly in terms of genericity, complexity of similarity measures, query language and visual analysis. The (also starting) PhD thesis of Guillaume Tejedor studies medical care sequences of patients suffering from Amyotrophic Lateral Sclerosis, aiming at the stratification of patients and the prediction of survival time [Tejedor et al., 2024].

Data narration. The primary objective of data exploration is to gain insights, which may then be contextualized, explained, structured and highlighted through varied visualizations in order to be communicated to an audience. This is a new paradigm, that of narrating stories with data [Marcel et al., 2023a].

The PhD thesis of Faten El Outa [Outa, 2023] proposes a conceptual model describing data narratives [Outa et al., 2020b] and a process model for data narration, which highlights the analysts' intentions [Outa et al., 2022, Outa et al., 2023]. The PhD thesis of Raymond Ondzigue Mbenga [Ondzigue Mbenga, 2023] specializes the data narration process to the field of public health, with an application to the epidemiological surveillance of tuberculosis in Gabon [Ondzigue Mbenga et al., 2022a]. Both theses draw attention to the importance of EDA support, the latter also dealing with data quality improvement.

We are currently interested in the automation of certain tasks [Chanson et al., 2022b], especially the derivation of queries from analysts' intentions [Francia et al., 2022c].

Data quality management. The study of data quality is omnipresent in my research activities since my PhD thesis [Peralta, 2006]. Quality issues complicate data exploration, influence analysts' behavior and can distort findings. My recent work places the data analysts (their contexts and analysis needs) at the core of data quality management.

The PhD thesis of Flavia Serra (defense planned for spring 2024) focuses on the fundamental aspects of data quality management. It models the components of the data context impacting quality [Serra et al., 2022a], and proposes a data quality management methodology that takes context into account [Serra et al., 2023]. The PhD thesis of Raphaël Bres (started in 2021), investigates the quality (data freshness) of geographic data, in particular cycling paths and lanes [Bres et al., 2023], and analyzes the impact on route recommendations [Bres et al., 2022]. We are currently interested in the modeling of the cycling network guided by data quality.

Remark that, regarding quadrants in Figure 1.1, themes (i) and (ii) mainly concern Data Exploration and Model Building and to some extents Exploitation, while theme (iv) concerns Data Engineering. Contrarily, theme (iii) covers all quadrants. This is sketched in Figure 1.2.

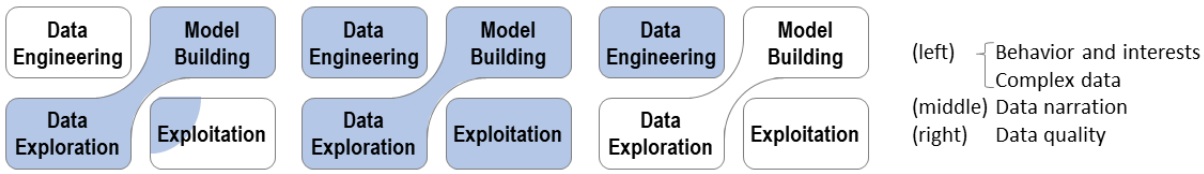


Figure 1.2: Coverage of research themes (blue shadow)

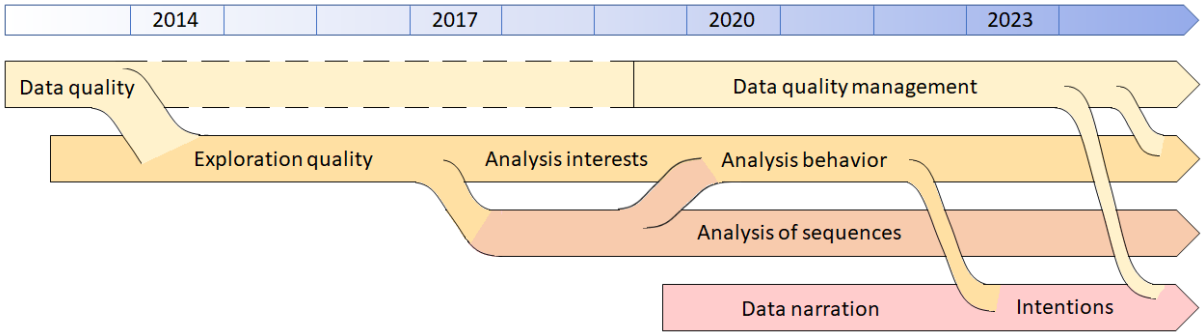


Figure 1.3: Timeline of research themes

As a timeline, Figure 1.3 situates my contributions in each theme and brings convergences out. Indeed, my background on data quality turned easily into exploration quality and indirectly influenced the other themes. Similarly, the analysis of sequences of queries, generalized to the analysis of semantic sequences, and in turn, our first results on clustering of semantic sequences enabled the study of analysis behavior. Later, this background on interests and behavior, enriched data narration with intentional aspects. Learned lessons on data quality management, also profit the other themes.

Table 1.1 lists my publications (in the 2014-2023 period) by theme, and Figure 1.4 summarizes my research experience in the form of a data narrative.

	Behavior and interests	**Complex data**	**Data Narration**	**Data Quality**
International Journals	[Djedaini et al., 2019] [Drushku et al., 2019] [Peralta et al., 2020] [Moreau et al., 2022] [Gkitsakis et al., 2024]		[Francia et al., 2022c] [Outa et al., 2023]	
International Conferences	[Aligon et al., 2014a] [Ba et al., 2014] [Furtado et al., 2015] [Djedaini et al., 2016] [Djedaini et al., 2017b] [Drushku et al., 2017] [Megasari et al., 2018] [Marcel et al., 2019] [Peralta et al., 2019b] [Chanson et al., 2020] [Moreau et al., 2020c] [Moreau and Peralta, 2021] [Gkitsakis et al., 2023]	[Moreau et al., 2021a] [Moreau et al., 2021b]	[Chédin et al., 2020] [Outa et al., 2020b] [Chanson et al., 2022b] [Ondzigue Mbenga et al., 2022a] [Outa et al., 2022] [Marcel et al., 2023a]	[Serra et al., 2022a] [Bres et al., 2023] [Serra et al., 2023]
National Conferences	[Boulil et al., 2014] [López et al., 2015] [Djedaini et al., 2017a] [Drushku et al., 2020]	[Tejedor et al., 2024]	[Ondzigue Mbenga et al., 2019] [Chagnoux et al., 2021] [Ondzigue Mbenga et al., 2021] [Outa et al., 2021] [Francia et al., 2022b] [Ondzigue Mbenga et al., 2022b]	[Bres et al., 2022]
Patents	[Drushku et al., 2021]			
Other Publications	[Peralta et al., 2019a] [Moreau et al., 2020a] [Gkitsakis et al., 2022]	[Moreau et al., 2020b]	[Outa et al., 2020a] [Peralta, 2020] [Marcel et al., 2023b] [Vassiliadis et al., 2024]	[Serra et al., 2022b]

Table 1.1: Recent publications by theme

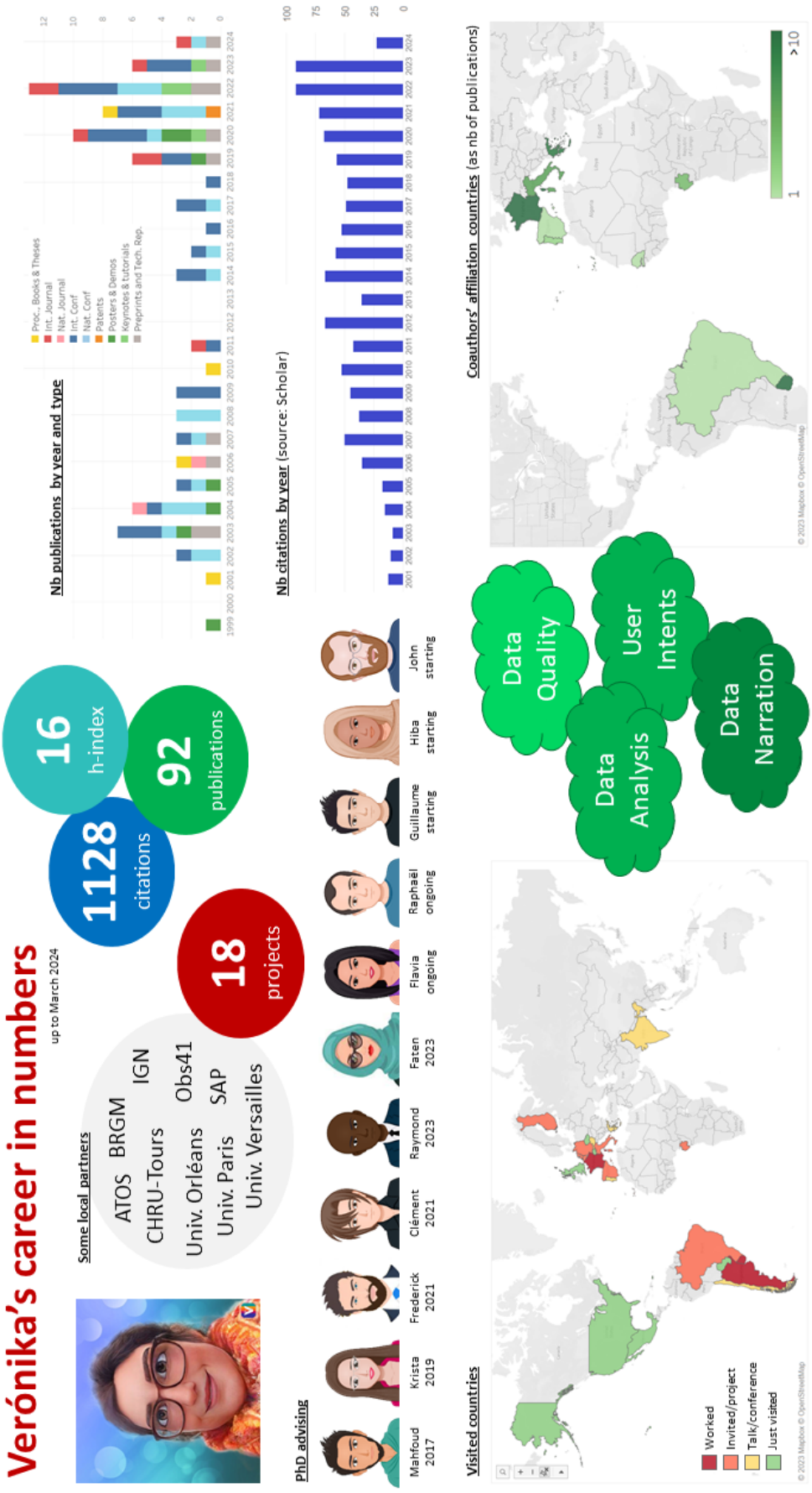


Figure 1.4: A data narrative summarizing my career

1.4 Selected contributions

This dissertation mainly focus on two of my research themes, namely, learning of analysts' behavior and interests –theme (i)– and data narration –theme (iii). Our contributions on analysis of complex data –theme (ii)– being tightly connected to theme (i), are regularly commented.

However, given that PhD theses on data quality management –theme (iv)– have not yet been defended, it seems more respectful to me to give them the honor of defending their proposals in front of a thesis jury. Thus, this dissertation does not comprehensively address their research work, but they are discussed at perspectives.

Consequently, this dissertation presents the following contributions:

Contributions

1. An approach for evaluating the quality of users' explorations, and indirectly, qualifying users' analysis skills,
2. A set of methods for segmenting a query workload into explorations, which is a mandatory first step for analysing publicly available query workloads lacking in metadata,
3. An approach for learning analysis patterns from users' explorations,
4. A comprehensive study of interestingness of a piece of data, rooted on the study of human behavior, and its usage for developing measurement algorithms,
5. An approach for learning users' interests in query workloads and leveraging them for query recommendation,
6. A conceptual model for data narratives, and
7. A process model for data narration, leveraging analysts' intentions.

1.5 Research projects and studied datasets

The contributions presented in this dissertation are based on data either from research projects where I participated, from user studies where I contributed, or from public datasets made available by their creators. This section introduces some datasets (query workloads and data narratives) used for validating the proposals (described respectively in Appendices A and B) and some research projects framing them (described in Appendix C). Table 1.2 relates them to our contributions.

1.5.1 Workloads

In what follows, we consider query workloads, typically arising from logs of database systems or query tools, containing (potentially long) sequences of queries made by some users. In particular, we focus on workloads of hand-written[6] queries. Most of them arise from experiments specifically designed to test an analysis tool or project, described in the literature. They all consist of navigation traces of real users on real data. Unlike them, some additional workloads were artificially generated using specialized generation tools.

[6]Consistently with the authors of [Jain et al., 2016], we use the term hand-written to mean, in this context, that the query is introduced manually by a human user, which reflects genuine interactive human activity over a dataset, with consideration between two consecutive queries.

In our experiments, we reuse such workloads for new tasks, as evaluating the quality of explorations, learning users' behavior and skills and discovering users' interests. Workloads, links, materials and usage are detailed in Appendix A.

Workloads of real users' queries:

- **Ipums** workload consists of navigation traces of students, collected during the testing phase of the development of Falseto, a tool meant to assist EDA [Aligon et al., 2014a].
- **Open** workload consists of navigation traces of students, collected in the context of the DOPAN project [Boulil et al., 2014], using Saiku OLAP tool.
- **Enterprise** workload consists in navigation traces of 14 volunteers of SAP company in the context of a research and innovation project [Drushku, 2019], using a SAP prototype that supports keyword-based BI queries.
- **Security** workload consists of analysis sessions made by expert analysts in the context of the Honeynet Project [Milo and Somech, 2018], using a prototype of web-based analysis platform.
- **SQLShare** workload is the result of a multi-year SQL-as-a-Service Experiment [Jain et al., 2016], allowing any user with minimal database experience to upload their datasets on-line and manipulate them via SQL queries.

Synthetic workloads:

- **Artificial** workload consists on artificial explorations generated using CubeLoad [Rizzi and Gallinucci, 2014], a tool for generating realistic explorations over star schemas, according to templates modeling various exploration patterns.
- **Loan** workload consists of several explorations over artificial data using a dedicated random generator [Gkitsakis et al., 2022]. This workload is used for scalability tests; the semantics of explorations is not exploited.
- **Adult** workload consists of few explorations, carefully (manually) devised to compare interesting aspects in a user study [Gkitsakis et al., 2022].

We chose to test our proposals in several workloads to avoid learning specific behavior of a set of users. Indeed, the considered workloads concern users with different analysis skills (students, novices, experts), using different analysis tools (open source tools, research prototypes, advanced user interfaces) and accessing datasets of different sizes and complexities. We are not aware of other public analytical workloads, specially from senior analysts, whose analysis activity is jealously guarded by companies as pointed out by [Rizzi and Gallinucci, 2014].

1.5.2 Data narratives

In what follows, we consider several data narratives, either crafted during user studies or publicly available. For some of them, we have also access to the data narration processes followed by the authors. See Appendix B for detailed description, visual snippets and links.

- **Narrating Rennes.** During the "Narrating Rennes by the data" challenge, 3 teams (among which journalists, students, social workers and data scientists) were observed during the crafting of a data narrative. The resulting data narratives take the form of a video, a notebook and an interactive book. A prize was awarded to the best one.
- **Fatal encounters**. For the "Fatal encounters" challenge, 24 teams of Master students specialized in data analysis, crafted data narratives after one-hour tutorial on data narration. Data narratives were assessed by an experienced data journalist, according to their quality and completion.

- **Strokes.** A data narrative informing women about stroke risks, published by GOOD company, in the form of an infographic.
- **Climate.** A data narrative about the climate crisis in the Sahel, published by OCHA United Nations office, in the form of a scrolly-story. The crafting process is documented in the blog of a data journalist
- **Tennis.** A data narrative about racket in tennis betting, published by BuzzFeed News, and its crafting process, documented by an investigative data reporter, both in the form of sport news.
- **Covid.** A data narrative about covid mortality in Alsace, published by Rue89 Strasbourg Newspaper in the form of a news article. The crafting process is documented by a data journalist in a notebook
- **Tuberculosis**. A data narrative about Tuberculosis pandemic in Gabon, developped and documented within the eGabonSIS project [Ondzigue Mbenga et al., 2022a].

1.5.3 Research projects

Several contributions are framed by the research projects listed below. See Appendix C for detailed description and links.

- **DOPAn** project focuses on the interactive analysis of open data, to study the energy vulnerability of households and territories. Its outcome is a user-friendly, user-centered dedicated Business Intelligence solution, the *Open* workload resulting from its testing phase.
- **Mobi'Kids** project studies the role of urban educative cultures in the evolution of children's daily mobility and life context. Its outcome is the collection and analysis of geolocated and semantically enriched tracks. Our methods for learning users' behavior were first experimented for mobility behavior.
- **Madona** is a cross-disciplinary project aiming at better understanding of the mechanisms of data selection and exploration that allow the gradual construction of data narratives. It envision the development of tools allowing journalists to explore open data with simplified interaction. The *Narrating Rennes* and *Fatal encounters* challenges were organized within this project. Project participants (specially a large panel of data journalists) tested our proposals.

As a summary, Table 1.2 indicates the workloads, data narratives and projects related to each contribution.

Contributions	Studied workloads and data narratives	Projects
1	Open, Enterprise	DOPAn
2	Open, Enterprise, SQLShare	
3	Ipums, Open, Security, Artificial	Mobi'Kids
4	Open, Loan, Adult	
5	Enterprise	
6	Strokes, Covid	Madona
7	Narrating Rennes, Fatal encounters, Climate, Tennis, Covid, Tuberculosis	Madona

Table 1.2: Summary of dataset usage and framing projects by contribution

1.6 Document organization

This dissertation has 3 chapters describing contributions and a final chapter providing conclusions.

Chapter 2 addresses the challenge of qualifying data explorations and learning analysis behavior from users' past explorations, represented in query workloads. It firstly introduces a model of queries and explorations from the prism of users' skills, based on a large set of features capturing various aspects of a query and its context within the exploration. Starting by a particular case of EDA, that is OLAP analysis of multidimensional data, exploration quality is learned in a simple, focused and rich environment. The proposal includes two classification models to evaluate to what extent a query is focused and contributes to the success of an exploration, and a knowledge tracing model to assess users' analysis skills (contribution 1). The extension from OLAP to a more complex SQL environment lacking in metadata introduced the challenge of workload segmentation. Three methods are proposed, comparing different strategies (contribution 2). Finally, the chapter describes an approach for clustering explorations revealing similar analysis patterns (contribution 3). The method is based on a similarity measure tailored for explorations, that assess whether explorations share similar sequences of operations and contain queries of close complexity.

Chapter 3 deals with users' interests and addresses the challenge of modeling and learning users' interests, improving users' analysis experience. It firstly study interestingness from the viewpoint of human behavior, and proposes a two-level framework for developing interestingness measures, consisting respectively of high-level interestingness aspects, and data-oriented assessment algorithms. Then, focusing in a particular interestingness aspect (the relevance of a query for the overall analysis intention of the user), the chapter describes an approach for learning users' interests in a query workload and recommending relevant queries. The discovery of coherent interests is formalized as a clustering problem, and a similarity measure is learned, intending to capture whether two queries reflect a same interest. To leverage the discovered interests for the purpose of query recommendation, an original interest-based recommender is proposed.

Chapter 4 considers EDA within the DN process and address the challenge of modeling the static and dynamic aspects of DN, setting the bases for the development of DN frameworks and tools. It describes a conceptual model for data narrative, providing a structured, principled definition of the key concepts of the domain, and a process model that covers the whole DN cycle and accommodates a wide range of practices observed in the field. This accommodation is evidenced by an instantiation of the models to the health domain and several use cases. Both models are backed by a large literature review and the observation of many (novice and expert) practitioners. Both models draw attention to the importance of EDA tasks and highlight intentional aspects.

Chapter 5 concludes this dissertation by reviewing and discussing the contributions, and introduces future research directions.

Finally, several **Appendices** provide detailed descriptions about some specific subjects. Specifically, Appendices A and B explain respectively the studied query workloads and data narratives. Appendix C describes past and current research projects. At last, Appendix D presents my list of publications and Appendix E my Curriculum Vitæ.

Chapter 2

Learning users' analysis behavior

This chapter describes our contributions for understanding, modeling and learning the way users analyse data.

It relies on materials published in several conferences and journals, the main ones being [Djedaini et al., 2019, Peralta et al., 2020, Moreau et al., 2022]. The overall contributions were developed in collaboration with several PhD and master students, as well as researchers and analysts of the DOPAn and Mobi'Kids projects, as summarized below.

Advising, projects and collaborations

PhD theses:
Mahfoud Djedaini (2014-2017), *Automatic assessment of OLAP exploration quality*, co-supervised with Patrick Marcel.
Clément Moreau (2018-2021), *Mining of semantic mobility sequences*[1], co-supervised with Thomas Devogele and Laurent Etienne.

Postdoctoral project:Kamal Boulil (2014-2015) co-supervised with Patrick Marcel.

Postgraduate project: Alexandre Chanson (2020).

Master theses and projects: Federico Mosquera (2015), Clément Chaussade (2017), Shibo Cheng (2017), Chiao Yun Li (2017), Martina Megasari (2017), Pandu Wicaksono (2017), Yann Raimond (2018), Willeme Verdeau (2018), Aboubakar Sidikhy Diakhaby (2019), Mohamed Ali Hamrouni (2019), Clément Legroux (2021), Mathis Rharbal (2021).

Research projects:
DOPAn – *Open data for monitoring and analysis*[2] (2014-2017), regional funding.
Mobi'Kids - *The role of urban educative cultures in the evolution of children's daily mobility and life context. Collection and analysis of geolocated and semantically enriched tracks*[3] (2017-2021), national funding (ANR).

[1]Written in French. Original title: *Fouille de séquences de mobilité sémantique*

[2]Original name (in French): *DOPAn - Données Ouvertes pour le Pilotage et l'Analyse*
`https://lifat.univ-tours.fr/lifat-english-version/projects/recent/bdtin/2014-2017-dopan`

[3]Original name (in French): *Mobi'Kids – Le rôle des cultures éducatives urbaines dans l'évolution des mobilités quotidiennes et des contextes de vie des enfants. Collecte et analyse de traces géolocalisées et enrichies sémantiquement*
`https://anr.fr/Projet-ANR-16-CE22-0009`

Contents

2.1 Problems and positioning

The analysis of a database workload to support EDA receives increasing interest from the database and machine learning communities (see e.g. [Idreos et al., 2015, Milo and Somech, 2020, Abadi et al., 2022, De Bie et al., 2022]) as it offers many practical interests, from the monitoring of database physical access structures [Chaudhuri and Narasayya, 2007] to the generation of user-tailored collaborative query recommendations for interactive exploration [Eirinaki et al., 2014, Milo and Somech, 2018].

Characterising users' behavior while analysing data, i.e. learning the way users analyse data (the type and order of operations, the complexity of queries, the level of detail, the degree of focus) is a step forward in the understanding of analysis activities and offers new applications.

The most natural one is a better support of EDA, for instance to understand users' information needs, to identify struggling during the exploration, or to provide better query recommendations. Notably, EDA systems usually do not offer such facilities. The prediction of next analysis steps is particularly interesting, enabling beforehand execution of probable queries and caching of results, as well as advanced optimization strategies.

Another benefit is the design of more realistic workloads for database benchmarking. Classical benchmarks like TPC-H or TPC-DS poorly include interactive exploration activities in their synthetic workloads, and are not appropriate to evaluate modern EDA systems [Eichmann et al., 2016]. Identifying analysis behavior would allow to better model user's explorations and mimic such activities in benchmark workloads.

Finally, we mention the detection of clandestine intentions as another potential benefit. Indeed, as reported by [Acar and Motro, 2004], query sequences may reflect such intentions, where users prefer to obtain information by means of sequences of smaller, less conspicuous queries to avoid direct queries which may disclose their true interests. The identification of typical analysis patterns may help distinguishing normal from clandestine intentions.

In what follows, we consider a *query workload*, typically arising from a log of a database system or query tool, containing a (potentially long) sequence of queries made by some users. In this context, a *session* is a raw sequence of queries (e.g. recorded during a user connection to a database system), while an *exploration* is a coherent sequence of queries, that all share the same goal of fulfilling a user's information need that may not be well defined initially.

Thus, explorations report on users' analysis activities, and therefore contain valuable raw material for studying users' analysis behavior. These topics have been studied for Web search from early 2000's [Mobasher, 2007] and interest many communities, in particular for the analysis of social networks [Abascal-Mena et al., 2015, Francia et al., 2019, Boukharouba et al., 2023].

Our research challenge is to **qualify data explorations and learn users' analysis behavior from users' past explorations, represented in a query workload**.

Several research needs arise. They are described in the following subsections.

2.1.1 Need for evaluation of exploration quality

Varied techniques have been proposed for supporting EDA, allowing users to interactively explore their data and help them to better gain insights. Nevertheless, there is yet no commonly agreed upon method for evaluating to what extent explorations conducted with such systems are indeed successful.

A first problem to investigate in this context is **how to evaluate the quality of users' explorations**, and indirectly, **how to qualify users' analysis skills**.

The database community enjoys a variety of popular benchmarks to assess and compare the performance of database systems. The TPC consortium[4] proposes benchmarks that include metrics covering time, performance, price, availability or energy consumption. However, while TPC acknowledges the importance of the explorative nature of decision support queries (see e.g., the OLAP interactive queries in the TPC-DS benchmark), none of the existing TPC metrics are appropriate for measuring database exploration support. In other words, the existing benchmarks adopt a system-centric viewpoint, measuring the efficiency of data retrieval, and are not appropriate to measure exploration efficacy under a user-centric angle.

Eichmann et al. also motivate the need for new, user-centric benchmarks and propose some tracks to investigate their building [Eichmann et al., 2016]. Considering that EDA main objective is to gain insights about the data, they propose the use of *number of insights per minute* as a primary metric for evaluating systems. They raise the challenges in defining such a metric, like defining user-specific insights and measuring the complexity of an insight.

Another interesting notion comes from Exploratory Search, a sub-domain of Information Retrieval that studies users' behavior during their explorations [White and Roth, 2009]. The basic model of exploration in Exploratory Search distinguishes two main phases. In a first phase, called *exploratory browsing*, users are likely to explore the space, as well as better defining and understanding their problem. At this stage, the problem is being limited, labeled, and a framework for the answer is defined. Over time, the problem becomes more clearly defined, and the user starts to conduct more targeted searches. In this second phase, called *focused phase*, users (re)formulate query statements, examine search results, extract and synthesize relevant information.

This notion of focus has not been previously studied in the database community. Nevertheless, detecting focused phases in data exploration can be exploited in a variety of applications, for instance in the context of data exploration assistants. When focused, an analyst would expect more precise queries, related to what she is currently analyzing. On the contrary, when exploring the data, the analyst would prefer more diverse queries, for a better data space coverage.

Beyond counting insights, our goal is to **characterize what makes an exploration successful and take advantage for evaluating exploration quality and users' skills**.

Our research track is to evaluate the degree to which a query contributes to the success of an exploration, in terms of user experience. We start by detecting focused queries and then turn to a more general notion of contributory queries. Intuitively, a query is contributory if it is related to an underlying information need, if it refines or generalizes previous queries, if it allows to investigate related data perspectives, if it returns new data not previously analyzed or allows to highlight unexpected data, briefly, if in some way it allows to increase user's knowledge about the studied phenomenon.

We remark that there is currently no formal and commonly agreed definition of query contribution, and that writing contributory queries can be seen as a form of procedural knowledge. Procedural knowledge is the knowledge about how to do something. Different from declarative knowledge, that is often verbalized, application of procedural knowledge may not be easily explained [Cauley, 1986]. However, models exist to automatically evaluate procedural knowledge acquisition [Corbett and Anderson, 1995].

We hypothesize that the procedural knowledge related to the skill of writing contributory queries can be modeled as a supervised machine learning problem, and investigate methods for learning a model of query contribution, and score the probability that the skill is mastered by the user.

Section 2.3 presents our contributions for assessing exploration quality.

[4]See http://www.tpc.org/ for details

2.1.2 Need for extraction of explorations

Query workloads typically consist of raw sequences of queries, made by some users, without further information on users' intentions and information needs. Indeed, explorations are not easily identifiable in a query workload. In the best case, queries are arranged in sessions, possibly containing several explorations.

An unmissable problem, specially in the case of large workloads, is to determine **whether a workload contains actual exploration activities**, and more particularly **how to extract such explorations**.

Session segmentation has been previously studied for the SDSS workload [Singh et al., 2007]. In their study, the authors consider that a new session starts after 30 minutes of think-time (time spent between two queries). A similar problem was largely studied for the segmentation of web traces (see for example [Wong et al., 2006]) proposing the same 30-minutes cutoff. Search engine providers, like MSN and Google, use similar heuristics.

Contrarily to those works, many workloads (e.g. SQLShare workload [Jain et al., 2016]) do not include query timestamps. Furthermore, even when timestamps are available, they may lead to wrong segmentation.

Our goal is to **segment a session in a smarter way**.

Our research track is to use machine learning methods for deciding whether to segment a session. We investigate three alternatives methods: (i) unsupervised learning, based only on similarity between contiguous queries, (ii) supervised learning, using transfer learning to reuse a model trained over a workload where ground truth is available, and, (iii) weak supervision, using weak labelling to predict the most probable segmentation from heuristics meant to label a training set.

Section 2.4 presents our contributions for workload segmentation.

2.1.3 Need for identification of analysis patterns

Once we are able to identify whether a workload contains actual exploration activities, the natural next challenge is to analyze such explorations in order to characterise users' analysis behavior.

Two main problems come out, **how to identify regular and unexpected analysis patterns**, and **what is the impact of query complexity**.

While some EDA support techniques (e.g. OLAP analysis) have been around for almost 30 years, little is known about typical navigational behavior. To the best of our knowledge, only two previous works relate to analysis patterns. The recurrent types of user analyses described in [Rizzi and Gallinucci, 2014] are the first attempt to define analysis patterns in OLAP workloads. Authors claim that obtaining real OLAP workloads by monitoring the queries actually issued in companies and organizations is hard, and propose a parametric generator of OLAP workloads, CubeLoad, based on four templates that model recurrent types of user analyses. In [Aligon et al., 2014a], we analysed a workload of explorations devised by master students and observe as general tendency that explorations are more focused and contain more relevant queries at the end.

Other works study query complexity in SQL logs. Jain et al. ran a number of tests on the SQLShare workload [Jain et al., 2016] showing the diversity and complexity of the workload. In [Vashistha and Jain, 2015], authors analyze the complexity of queries in the SQLShare workload, in terms of some query features (e.g. number of tables, columns, characters and operators) and

query run-time. They define two complexity metrics from these features: the Halstead measure (traditionally used to measure programs complexity) and a linear combination of query features learned using regression.

Our goal is to go a step forward and **learn more analysis patterns from the explorations of real users**. Concretely, we aim to cluster together explorations showing similar analysis patterns. The idea behind analysis patterns is to look for sequences of common operations performed together when analysing data, as some kind of movements in a data space. Query complexity, both in terms of expressiveness and usage of advanced clauses, may be a good indicator of analysis behavior, complementing the study of operations.

Our research track is to cluster together similar explorations, sharing similar sequences of operations and containing queries of close complexity. To this end, we investigate similarity measures and clustering algorithms tailored for explorations.

Section 2.5 present our contributions for learning analysis patterns and query complexity.

2.1.4 Scope

We start by studying a particular case of EDA, OLAP analysis of multidimensional data, before tackling the regular case of analysis of relational data.

We choose to first focus on multidimensional data organized in cubes due to (a) their extreme relevance to the problem, as analysts explore data in query sessions via Business Intelligence tools, (b) their simplicity, as the simplest possible database setting in terms of how data are presented to the end-users, (c) their most focused setup, also due to the simplicity of the underlying schema, but also because the queries follow a pattern of filtering and grouping with very specific joins between dimension and fact tables, and, (d) the richness of information content, due to the presence of hierarchically structured dimensions that allow manipulating, examining and understanding the data from multiple layers of abstraction. This last property is also what differentiates cube queries from regular, relational ones: the presence of a hierarchical multidimensional space allows comparisons at multiple levels of granularity that would otherwise be very hard to express or detect in a plain relational environment.

Therefore, the analysis of OLAP workloads allows to qualify explorations and learn users' analysis behavior in a relevant, simple, focused and rich environment; the simile is like solving the problem in vitro in a lab, before addressing it in an industrial factory.

Transposing such approach to regular, non multidimensional SQL workloads raises many challenges. Even if a sequence of SQL queries is issued to explore the database content, non multidimensional relational schemata do not have the regularities one expects from the multidimensional model, explorations may not be expressed through roll-up or drill-down operations, SQL queries may deviate from the traditional star-join pattern commonly used for analytical purpose, etc. In addition, hand-written SQL queries may be of varied complexity, in comparison to the typical star-join queries frequently generated by OLAP tools.

In this way, the analysis of SQL workloads offers the opportunity to learn users' skills, translated by the type of operations, functions and clauses used in SQL queries.

Road map. Section 2.2 introduces our representation of queries and explorations. Then, Sections 2.3, 2.4 and 2.5 present our contributions to the previously described research needs and Section 2.6 draw our conclusions.

2.2 Query and exploration models

This section introduces our representation of database queries and explorations. We use the term *query* to denote the text of a hand-written query statement, and *exploration* to denote a coherent sequence of queries, that all share the same goal of fulfilling a user's information need that may not be well defined initially.

For each query, we extract a set of *fragments*, such as projections, selections and aggregations, that abstract the most descriptive parts of a query. Then, we compute a set of *features* representing main characteristics of the query itself (e.g. the number of selections), its relationship with previous query in the exploration (e.g. the number of common selections), and its relationship with the whole operation (e.g. its position in the exploration).

We describe here a large subset of features, allowing the description of many types of queries, from simple star-join queries to SQL queries of arbitrary complexity. The learning tasks described in this manuscript are based on subsets of such query features.

Basic knowledge is assumed on the relational model and query languages, as can be found in e.g., [Abiteboul et al., 1995], and on BI models and query languages, as described in e.g., [Golfarelli and Rizzi, 2009].

2.2.1 Query fragments

We represent queries as a collection of fragments extracted from the query text, such as projections, selections, aggregations, tables, group by and order by expressions. These fragments abstract the most descriptive parts of a query, and are the most used in the literature (see e.g., [Khoussainova et al., 2010, Eirinaki et al., 2014, Milo and Somech, 2018]). We also consider the overall attributes explicitly appearing in the query (which informs about user's effort in writing the query) and more complex fragments, namely, sub-queries, functions and some complex clauses, that even less frequently used, indicate query complexity. A quantitative analysis of queries in the SQLShare workload [Jain et al., 2016] motivates the choice of such complex fragments.

Definition 2.1 (Query) *A query representation, or with a slight abuse of language, a* ***query****, over relational database schema DB is a 11-uple* $q = \langle text, P, S, A, T, G, O, At, Sq, F, C \rangle$ *where text is the full text of the query, and P, S, A, T, G, O, At, Sq, F and C are the sets of query fragments, resp. projections, selections, aggregations, tables, group by expressions, order by expressions, attributes*[5]*, subqueries, named functions and complex clauses*[6]*.* □

We intentionally remain independent of presentation and optimization aspects, specially the order in which attributes are projected (and visualized by the user), the order in which tables are joined, etc. All the queries we consider are supposed to be well formed, and so we do not deal with query errors.

Finally, an **exploration** is a sequence of queries of a user over a given database. In addition, when query execution timestamps are available, we consider the timestamps before and after the execution of each query.

[5] Attributes appearing explicitly in the query. Expressions, views, sub-queries and other clauses are parsed in order to obtain the referenced attributes. This allows to consider all attributes, even those that are part of atypical or less-frequently-used clauses.

[6] Other advanced and expert clauses. For example, from a quantitative analysis of clauses used in the SQLShare workload, we selected the following ones: TOP, HAVING, CASE, LEFT OUTER JOIN, RIGHT OUTER JOIN, INNER JOIN, FULL JOIN, UNION, EXCEPT, INTERSECT, PIVOT and OVER.

Query language	**Workload**	**Tool producing the workload**	**Parser description and implementation**
SQL	SQLShare	Microsoft SQL Server	https://github.com/Belisaire/stage
MDX	Open	Saiku	[Djedaini, 2017] http://github.com/mdjedaini/ideb/
OLAP-like	Artificial	CubeLoad [Rizzi and Gallinucci, 2014]	
	Ipums	Falseto [Aligon et al., 2014a]	
	Enterprise	SAP BI prototype [Drushku et al., 2019]	Omitted for confidentiality
	Security	Analysis prototype [Milo and Somech, 2018]	[Moreau et al., 2020c] https://github.com/ClementMoreau-UnivTours/CED_Dolap

Table 2.1: Parsers developed for several workloads

Definition 2.2 (Exploration) *Let DB be a database schema. An exploration $e = \langle q_1, \ldots, q_p \rangle$ over DB is a sequence of queries over DB. We note $q \in e$ if a query q appears in the exploration e, and exploration(q) to refer to the exploration where q appears.* □

Several parsers were implemented for extracting query fragments from query workloads, each one specialized for a particular query dialect or log format. Table 2.1 summarizes them and redirect to their detailed descriptions.

2.2.2 Query features

For each query q_k in an exploration e, we compute a set of features that intend to capture different aspects of the query and its context. For the sake of presentation, we categorize features as follows: i) intrinsic features, i.e., only related to the query itself, ii) relative features, i.e., also related to the query's predecessor in the exploration, and iii) contextual features, i.e., related to the whole exploration, providing more context to the query. They are listed[7] respectively in Tables 2.2, 2.3 and 2.4.

Intrinsic features can be computed only considering the query q_k, independently of the exploration e and other queries in e. They intend to quantify many facets of the query, namely, (i) its analytical parts, in terms of level of aggregation, filters and measures, which are the main components of analytical queries (see e.g. [Golfarelli and Rizzi, 2009, Aligon et al., 2015]), (ii) its complexity, captured through the length of the query (in terms of written characters, attributes, tables, functions and sub-queries, and thus measuring user's effort to write the query), the usage of advanced clauses (informing on user's level of expertise) and execution time, these criteria being inspired from previous studies (see e.g. [Vashistha and Jain, 2015]) or emerging from preliminary observations of users' behavior[8], and (iii) the richness of query answers, in terms of size and contained information, both criteria largely used in the domains of query personalization and information retrieval (see e.g. [Bellatreche et al., 2005, Belkin et al., 2003]).

Relative features are computed comparing the query q_k to the previous query q_{k-1} in the exploration e. They capture both, (i) their commonness and differences (closer queries revealing more focused analysis, as studied in exploratory search ([White and Roth, 2009]), and (ii) the type of analytical operations that connect queries (i.e. that express a query w.r.t. the previous

[7]We remark that some features are referred with several names in our contributions, sometimes because their usage has evolved or simply for abbreviation. In these cases, all names are listed in the respective tables, for facilitating the link with the cited articles.

[8]For example, beginners and expert users, both can write focused queries, but the type of clauses may significantly differ.

Names	Description	Usage
P, NoP	Number of projections	seg
S, NoS, NoF	Number of selections (filtering predicates)	seg, ctr, foc
A, NoA, NoM	Number of aggregations (measures in OLAP)	seg, ctr, foc
T, NoT	Number of tables	seg, cmp
G, NoL	Number of expressions (levels in OLAP) in the group-by set)	ctr, foc
ADepth, LDepth	Aggregation depth (sum of depths of levels in the group-by set)	ctr, foc
FDepth	Filter depth (sum of depths of levels appearing in selections)	ctr, foc
C, NoCh	Number of characters	seg, cmp
B, NoAt	Number of attributes	seg, cmp
Q	Number of sub-queries	cmp
F	Number of functions	cmp
J	Number of advanced join types	cmp
U	Number of set operator types	cmp
V	Number of advanced clause types	cmp
E	Number of expert clause types	cmp
NoC	Number of cells (in query answer)	ctr, foc
QoI, RNI	Quantity of information (contained in query answer; relevant new info.)	ctr, foc
ExecTime	Execution time	ctr, foc

Table 2.2: Intrinsec query features and their usage for learning: exploration segmentation (seg), query contribution to exploration quality (ctr), query focus (foc) and query complexity (cmp)

Names	Description	Usage
NCP	Number of common projections	seg
NCS, NCF	Number of common selections (filtering predicates)	seg, ctr
NCA, NCM	Number of common aggregations (measures in OLAP)	seg, ctr
NCT	Number of common tables	seg
NCL	Number of common levels (in the group by set)	ctr
RED, IED	Relative edit distance (effort to express a query starting from the previous one)	seg, ctr, foc
JI	Jaccard index (of common query fragments)	seg
RI	Relative identity (Whether the query is identical to the previous one)	ctr
RR, IR	Relative recall (recall of cells in query answer w.r.t. previous answer)	ctr, foc
RP, IP	Relative precision (precision of cells in query answer w.r.t. previous answer)	ctr, foc
IsRefine	Is refinement (whether the query is a refinement of the previous one)	ctr
IsRelax	Is relaxation (whether the query is a relaxation of the previous one)	ctr
+P	Number of added projections	pat
-P	Number of deleted projections	pat
+S, NAF	Number of added selections (filtering predicates)	pat
-S, NDF	Number of deleted selections (filtering predicates)	pat
+A, NAM	Number of added aggregations (measures in OLAP)	pat
-A, NDM	Number of deleted aggregations (measures in OLAP)	pat
+T	Number of added tables	pat
-T	Number of deleted tables	pat
+G, NAF	Number of added group by expressions (levels in OLAP) in the group by set	pat
-G, NDL	Number of deleted group by expressions (levels in OLAP) in the group by set	pat
+O	Number of added order by expr.	pat
-O	Number of deleted order by expr.	pat

Table 2.3: Relative query features and their usage for learning: exploration segmentation (seg), query contribution to exploration quality (ctr), query focus (foc), and analysis patterns (pat)

Names	Description	Usage
CPQ	Click per query (number of successor queries that differ in at most one operation)	ctr, foc
ClickDepth	Length of the sequence of successor queries that differ in at most one operation	ctr, foc
IVA	Increase in view area (number of cells in query answer that were not seen previously)	ctr, foc
NoQ	Number of queries (executed so far, i.e. absolute position of the query in the exploration)	ctr, foc
QRP	Query relative position	ctr, foc
ElapsedTime, ElTime	Elapsed time since the beginning of the exploration)	ctr, foc
QF	Query frequency (number of queries executed so far, per unit of time)	ctr, foc
ConsTime	Consideration time (spent in analyzing query answer)	ctr, foc

Table 2.4: Contextual query featuresand their usage for learning: query contribution to exploration quality (ctr), and query focus (foc)

one, for instance projecting additional attributes, drilling down), which informs about user's habits (specially through the most used operations), and to some extent user's level of expertise (through how varied and complex are such operations).

Contextual features are exploration-dependent and make sense only in the context of an exploration. The same query q_k occurring in different explorations may be given different scores for features in this category. They capture both, (i) the place of the query in the exploration (in terms of position in the sequence, but also of elapsed time from the beggining of the exploration), and, (ii) the commonness and differences with other queries in the exploration (for instance, through the number of successor queries that differ in at most 1 operation, or the number of cells in query answer that were not seen previously).

Features are computed from query fragments. In particular, for intrinsic features, we consider a query $q_k = \langle text_k, P_k, S_k, A_k, T_k, G_k, O_k, At_k, Sq_k, F_k, C_k \rangle$, occurring at position $k \geq 1$, in the exploration e over the instance I of schema DB. Relative features also consider the previous query in the exploration, $q_{k-1} = \langle text_{k-1}, P_{k-1}, S_{k-1}, A_{k-1}, T_{k-1}, G_{k-1}, O_{k-1}, At_{k-1}, Sq_{k-1}, F_{k-1}, C_{k-1} \rangle$. For the particular case of the first query of e, i.e. q_1, we consider as predecessor the "empty" query $q_0 = \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle$. Contextual features consider all queries in the exploration.

We remark that the computation of some features need the execution of the query (e.g. for measuring execution time or counting the cells in query answer), thus depending on the availability of users' datasets, and other features relay on the existence of timestamps in the workload (e.g. elapsed and consideration time). As such information is not available for all the workloads, in some cases, several features need to be estimated or ignored, sometimes limiting to the subset of features that may be computed exclusively from query text. For instance, in the SQLShare workload, only a portion of users' datasets are available for confidentiality reasons (i.e. many users did not agree to share their data), and there are no timestamps.

We also remark that some learning tasks do not need the complete set of the features, for example, the evaluation of query complexity only relies on intrinsic features, not depending on other queries of the exploration. The learning tasks where features are used (which are described in the following sections and chapters) are mentioned in the usage column of Tables 2.2, 2.3 and 2.4.

2.2.3 Query vectors

In what follows, we represent a query in the space of query features, i.e. as multidimensional vectors, each position corresponding to one of the features described in Tables 2.2, 2.3 and 2.4. We consider several types of vectors, grouping the features relevant to specific learning tasks, for example, those concerning analytical operations or query complexity. This representation is at the core of our proposal for computing the similarity between queries, delimiting explorations and learning users' behavior.

Definition 2.3 (Query vector) *Let q_k be a query and q_{k-1} its predecessor in an exploration e. Let $F = \langle F_1, ...F_m \rangle$ be a sequence of features. A **query vector** is a m-dimensional vector $v = \langle v_1, ...v_m \rangle$ where $v_i = F_i(q_k, q_{k-1}, e)$ for $1 \leq i \leq m$.* □

Example 2.1 *Consider an exploration e_1 composed of 4 queries:*
q_1: SELECT species FROM All3col;
q_2: SELECT species FROM All3col WHERE longitude < 0;
q_3: SELECT species, longitude, latitude FROM All3col;
q_4: SELECT species, longitude FROM All3col ORDER BY species;
and consider a sequence of features $\langle +P, -P, +S, -S, +T, -T \rangle$, counting the variations (added/deleted) in projections, selections and tables.

Query vector for q_1, $\langle 1, 0, 0, 0, 1, 0 \rangle$, indicates an added projection (species) and an added table (All3col) w.r.t. the empty query.

Query vectors for q_2, q_3 and q_4, $\langle 0, 0, 1, 0, 0, 0 \rangle$, $\langle 2, 0, 0, 1, 0, 0 \rangle$ and $\langle 0, 1, 0, 0, 0, 0 \rangle$ resp., indicate the differences w.r.t. previous queries, i.e., an added selection (longitude < 0), 2 added projections (longitude, latitude) with a deleted selection, and 1 deleted projection. □

As query vectors may be long and have many 0-valued coordinates, we concisely represent them by listing the occurring operations (the ones not 0-valued) in the form "$\pm nX$", where X is a feature name[9] and $n \geq 1$ is its magnitude (omitted if 1). Signs are used only for some relative features (e.g. +P, -P), and are placed before magnitude for ease of lecture. For instance, the vectors of queries of Example 2.1 can be noted +P+T, +S, +2P-S and -P, respectively.

Finally, in some analyses in next sections, we focus on the presence of a fragment (e.g. projections), disregarding the magnitude (e.g. how many projections are concerned) and sign (addition or deletion). To this end, we compute aggregated vectors, as Boolean vectors with one dimension per concerned fragment. Analogously, they can be concisely represented using feature names (e.g. P,S,A,T,G,O). For instance, the aggregated vectors of queries of Example 2.1 can be noted PT, S, PS and P, respectively.

We remark that some features are expected to have disparate magnitude. Indeed, the number of characters is expected to be significantly higher than the number of attributes and the latter than other features as the number of tables, functions and sub-queries. Consequently, features C and B may need normalization; parameters will be determined experimentally (see next sections). Feature selection may also be necessary, especially if some features are highly correlated. The choice and setting of features is also discussed in next sections.

In next sections we use this representation of queries and explorations for mining exploration quality, segmenting query workloads and learning analysis patterns.

[9]This representation is practical only for features with abbreviated names, typically referring to query fragments, e.g. P, S, A, T, G, O.

2.3 Mining exploration quality

Foreword

This section summarizes part of the PhD thesis of Mahfoud Djedaini, co-supervised with Patrick Marcel. It also concerns the master projects of Clément Chaussade (2017), Shibo Cheng (2017), Chiao Yun Li (2017), Martina Megasari (2017) and Pandu Wicaksono (2017), co-supervised with Patrick Marcel and Nicolas Labroche.
The proposal was published at ADBIS [Djedaini et al., 2017b] and extended at Information Systems [Djedaini et al., 2019].

In this section we present an approach for evaluating the quality of users' explorations, and indirectly, qualifying users' analysis skills. We place in OLAP context, which represents a relevant, simple, focused and rich environment for tackling the problem.

We investigate three tracks, using supervised machine learning methods:

- query focus: our first method learns an interpretable model of query focus as a classification problem, based on query features,
- query contribution: our second method extends the first one for learning a model of query contribution,
- users' skills: our third method scores users' analysis skills as a procedural knowledge acquisition problem.

While there exists no formal definition or consensual formula to decide whether OLAP explorations and queries are focused or contributory, these concepts can be intuitively described by different characteristics that indicate a focused and high-quality activity. Our hypothesis is indeed that the definitions of focus and contribution are highly dependent of a fine characterization of the queries composing an exploration. For instance, the granularity level or the number of filters of a query, or the number of OLAP operations that separate two consecutive queries, are such characteristics.

We also hypothesize that the skill of *writing contributory queries* can be modeled as procedural knowledge and used for scoring the probability that a user masters the skill.

Considered workloads. We experiment on two workloads of real explorations, Open and Enterprise, described in Appendix A.

In order to build a ground truth, explorations were labeled by experts, using a labeling tool specifically designed for that purpose. They used Boolean labels for focus and for contribution. In addition, explorations were also manually inspected by an expert (a lecturer) and tagged with A-B-C labels, according to their overall quality. Label A corresponds to skilled users devising good explorations, label B to users that are learning analysis skills but still produce middle-quality explorations, and label C to low-skilled users devising poor explorations.

Query features. We investigate the impact of a large number of features, to finely describe different aspects of a query, either intrinsically, relatively to its predecessor query or relatively to the whole exploration containing it. Concretely, we used 19 query features for learning query focus (those tagged *foc* in Tables 2.2, 2.3 and 2.4). This set was extended with 6 additional relative features for learning query contribution (totting up 25 query features, those tagged *ctr* in the same tables).

We remark the absence of timestamps in the Enterprise workload, which perturbs (or prevent) the computation of some query features. In particular, we re-executed the queries in order to compute execution time (ExecTime), but we have no information about consideration time (ConsTime), which had to be excluded from the model. Finally note that elapsed time (ElapsedTime) is only computed on the basis of execution time.

Next subsections describe the three proposed methods.

2.3.1 Query quality

In order to learn query quality, we first propose a model of query focus, which is then extended to the more general model of query contribution.

Learning query focus Our first method aims at automatically detecting focus phases in users' explorations. As mentioned above, there is yet no formula for deciding whether a query is focused or not. However, an expert is able to recognize a focus activity by looking at various characteristics of the queries and the exploration.

In order to quantify these intuitive characteristics, we define a set of features, which characterize different aspects of a query: the user intention (e.g., the desired granularity expressed through the aggregation level), the results (e.g., the number of cube cells retrieved), as well as its relationship to other queries (e.g., the differences between a query and its predecessor).

Then, the problem of formally characterizing a focused query can be expressed as a classification problem in which a query is represented by a query vector and the class output variables is binary, either "focused" or "not focused". These are the only two classes we are able to define regarding the fuzzy notion of focus.

As the ability to interpret what makes a query focused is a major objective in our work, we limit ourselves to linear models that learn a weight for each query feature and then output a focus score that is computed as a weighted sum over the features for each query. In this context, we use an off-the-shelve SVM classifier whose separative hyperplane equation provides the expected relation to qualify the focus of a query based on our features and their associated weights. Moreover, this formalization allows to understand in a very intuitive way how each feature contributes to the detection of focus.

Learning query contribution Our second method generalizes the previous one, aiming to learn to what extent queries contribute to a successful exploration.

Intuitively, a query is contributory if in some way it allows to increase user's knowledge and gain insights.

Consequently, contributory queries may:

- relate to the information need guiding the exploration,
- refine or generalize previous queries,
- investigate related data perspectives,
- return new data not previously analyzed,
- highlight unexpected data.

As in previous method, we represent queries using query features (actually we extend the set of query features) and we formalize the problem as a binary classification problem. We choose a linear SVM classifier for the same reasons.

2.3.2 Exploration quality

Our third method gives an overall score to an exploration that corresponds to the probability that the skill of writing contributory queries is mastered by the analyst.

We use a classical model of skill acquisition, called Bayesian Knowledge Tracing (KT) [Corbett and Anderson, 1995], that estimates the probability that a skill is mastered from a collection of opportunities to use the skill. In our context, each query corresponds to an opportunity to contribute to the exploration.

We first introduce the basics of KT, and then describe our approach.

Bayesian Knowledge Tracing As mentioned before, procedural knowledge is the knowledge about how to do something, which application may not be easily explained [Cauley, 1986]. Many models exist to evaluate procedural knowledge acquisition. One of the most popular and successful models is Bayesian Knowledge Tracing [Corbett and Anderson, 1995]. An individual's grasp of the procedural knowledge is expressed as a binary variable, L, expressing whether the corresponding skill has been mastered or not. The knowledge of an individual cannot be directly observed, but can be induced by the individual answering a series of questions (or opportunities to exercise the skill) to guess the probability distribution of knowledge mastering. Measuring the skill mastery is noted $P(L_i)$, which corresponds to the probability that the skill L is mastered after answering i questions. Observation variables, X_i, are also binary: the answer to the question is either correct and wrong.

Specifically, the Knowledge Tracing model has four parameters, namely, two learning parameters, $P(L_0)$ and $P(T)$, and two performance parameters, $P(G)$ and $P(S)$. $P(L_0)$ is the probability that the skill has been mastered before answering the questions. $P(T)$ is the knowledge transformation probability: the probability that the skill will be learned at each opportunity to use the skill (i.e., the transition from not mastered to mastered). $P(G)$ is the probability of guessing: in the case of knowledge not mastered, the probability that the individual can still answer correctly. $P(S)$ is the probability to slip, i.e., to fail while the skill is already mastered. The model uses these parameters to calculate the learning probability after each question to monitor individual's knowledge status and predict their future learning probability of knowledge acquisition using a Bayesian Network.

Hawkins et al. proposed a fitting method that allows the empirically derivation of the four parameters [Hawkins et al., 2014]. Wang et al. proposed to extend the Knowledge Tracing model by replacing the discrete binary performance node with continuous partial credit node [Wang and Heffernan, 2013]. These two improvements of the Knowledge Tracing model (in the fitting method and the use of partial credits) were used successfully in sequencing educational content to students [David et al., 2016].

Assessing the overall quality of an exploration The contribution model presented in Subsection 2.3.1 allows to give a contribution score to each query of an exploration. Then, each exploration can then be seen as a sequence of scores, for each of its queries. In this way, each step of an exploration can be treated as an opportunity to exercise the skill of writing contributory queries.

Therefore, a KT model can be used, based on these contribution scores, to predict the skill of writing contributory queries for a specific user. To do so, as our $contrib(q)$ scores are real-valued, we use the extension of the KT model to continuous partial credits [Wang and Heffernan, 2013], which has been proven to evaluate skills more precisely than the binary KT. In this extension, $P(G)$ and $P(S)$ are assumed to follow a Gaussian distribution, and as such, these two quantities are represented by a mean value and a standard deviation. As a consequence, and as opposed to

the binary KT, the prediction $P(L_n)$ also follows a Gaussian distribution, whose mean is used as the value of the prediction and whose standard deviation expresses the confidence attached to this prediction.

To learn the 6 parameters of the continuous KT, we extend the approach proposed in [Hawkins et al., 2014] so that it outputs estimates of $P(G)$ and $P(S)$ described by a mean and a standard deviation. Then, based on these 6 parameters, the estimation of each skill acquisition $P(L_n)$ is performed by running 100 tests with each time randomly generated values for $P(G)$ and $P(S)$ following their respective distribution. From these 100 $P(L_n)$ estimates, we compute a mean and a standard deviation following the normal hypothesis. In the end, the mean $P(L_n)$ is the overall score of the exploration and the standard deviation is the confidence in this prediction.

It is important to note that we apply the KT on each exploration independently, even if the KT parameters are learned from a representative set of explorations.

2.3.3 Experiments and results

In this section we report the major findings of our experiments for qualifying explorations and users' skills.

In what follows, we consider two workloads with ground truth: Open and Enterprise. The technical differences of such workloads, as well as their different types of users and information needs, provide a good opportunity for testing our approach in different configurations.

Protocol. Query workloads are preprocessed for extracting query fragments and computing query features, as described in Section 2.2. Query features are normalized to avoid bias in the interpretation of model weights.

Then, a focus model is trained on the labeled queries of the Open workload, and two contribution models are trained on the labeled queries of the Open and Enterprise workloads, respectively. We use a linear SVM classifier with oversampling and 10-fold cross validation; it outputs coefficients that traduce the relative importance of each feature.

In our first experiment, we interpret the weights and measure the accuracy of each model to assess to what extent the learned models are consistent with the human expertise. The use of several workloads aims to investigate whether contribution models are sensible to the application context. Indeed, workloads have different types of users (students vs. analysts); different types of information needs (fuzzy vs. predefined); different characteristics of underlying data (large cubes with tens of dimensions and measures vs. small cubes with specific data); different query tools (Saiku vs. SAP prototype). Another important difference is that queries in each dataset were labeled by different experts. This is a major issue as the concept of query contribution (as the notion of quality itself) is fuzzy and highly dependent on the evaluator.

The second experiment aims to investigate if a model learned on a workload can be generalized to other workloads without any significant loss in prediction rate. To this end, we train a model on the Open workload (the whole set of queries) and test it on the Enterprise workload (also all queries), and the other way around. It is expected to obtain worse results, because of all the differences between datasets and annotation evoked below.

The objective of our last experiments is to verify our primary hypothesis assuming that skilled users are more likely to develop better explorations. To this end, we firstly compute the focus (resp. contribution) of an exploration as the average of the focus (resp. contribution) of its queries, and we compare to experts' labels on user's skills. Then, we score explorations using our continuous Knowledge Tracing (KT hereafter) prediction model, and we also compare to experts' labels on user's skills.

Implementation and setting. All experiments are run on a 64 bits Windows 8.1 Operating System, featuring a Intel(R) Xeon(R) CPU E3-1241 v3 @3.50GHZ and 16GB of RAM. Our prototype is written in Java 8 and Python 3, with Scikit-learn and Imbalanced-learn [Lemaitre et al., 2017] packages. It can be downloaded from GitHub[10].

Query features are normalized using z-score. As we observe acceptable levels of correlation and the models with the whole set of features provides higher accuracy, we keep all the considered features in the experiments hereafter.

We compare several strategies to balance the 2 classes of our datasets, either by over-sampling the minority class or under-sampling the majority class. Even if all results are very close in terms of accuracy, precision and recall, we choose ADASYN oversampling technique, which provides the best results and provides a slightly larger dataset after resampling.

Learned models. The obtained focus and contribution models are presented in Tables 2.5 and 2.6, resp. The impact of each feature can be positive/negative in terms of polarity, and high/low in terms of intensity. Here, we highlight trends and discuss some of the features.

Focus model. A first general observation is that all categories of features are important, as they include features having high weights. This means that the query itself, but also its context, indeed provide semantics when assessing focus.

A focused analyst has a relatively well defined information need in mind, which is clearly evidenced by the weights discovered. Indeed, among the features related to query text and answers, we observe that all the features that restrict the perimeter of the analyzed data (like NoM, NoF, NoL, ADepth, FDepth) have a positive impact on focus. And as expected, features that relax the perimeter of analyzed data (like NoC and IVA) appear to have a negative impact on focus.

Also, as expected, features that measure the closeness between two consecutive queries have a positive impact on the focus. RP is the best representative of that in the sense that its value decreases with the number of new cells gathered compared to cells in the previous query. Contrarily, features that characterize an important move within the data space (like RED and IVA) have a negative impact on focus.

Interestingly, most features relative to chronology have little impact on focus, with the notable exception of NoQ, which tends to confirm that focus phases indeed happen after rather long exploratory phases. Another rather surprising finding is that complex features, like QoI (quantity of informantion) do not show a significant impact on focus.

Intrinsic features		**Relative features**		**Contextual features**	
NoM	0.246	RED	-0.201	CpQ	-0.100
NoF	0.553	RR	0.008	ClickDepth	0.491
NoL	0.192	RP	0.203	IVA	-0.051
ADepth	0.217			NoQ	0.176
FDepth	0.147			QRP	-0.057
NoC	-0.395			QF	0.019
QoI	0.068			ElapsedTime	0.007
ExecTime	0.030			ConsTime	0.084

Table 2.5: Model of query focus (features and their weights) on the Open workload

[10]http://github.com/mdjedaini/ideb/

Intrinsic features		Relative features		Contextual features	
NoM	0.294	NCM	-0.167	CpQ	0.043
NoL	0.289	NCL	0.243	ClickDepth	0.247
NoF	0.406	NCF	0.241	IVA	0.032
ADepth	-0.133	RED	-0.025	NoQ	-0.057
FDepth	-0.431	RI	0.568	QRP	-0.099
NoC	-0.117	RR	0.108	QF	0.006
QoI	-0.021	RP	0.184	ElapsedTime	-0.211
ExecTime	-0.235	IsRefine	0.421	ConsTime	0.142
		IsRelax	0.174		
Bias $w_0 = 0.076$					

Intrinsic features		Relative features		Contextual features	
NoM	0.034	NCM	0.151	CpQ	-0.015
NoL	0.604	NCL	-0.009	ClickDepth	0.042
NoF	0.001	NCF	-0.070	IVA	-0.086
ADepth	0.569	RED	-0.025	NoQ	0.027
FDepth	-0.109	RI	-0.694	QRP	-0.890
NoC	0.000	RR	-0.621	QF	-0.038
QoI	1.130	RP	-0.399	ElapsedTime	0.133
ExecTime	0.004	IsRefine	0.118		
		IsRelax	-0.141		
Bias $w_0 = 0.364$					

Table 2.6: Models of query contribution (features and their weights) on the Open (top) Enterprise (bottom) workloads

Contribution models. A first general observation is that the importance of the features is substantially different in the models obtained on the Open and Enterprise workloads, as evidenced in Table 2.6. In particular, 15 out of 24 common features have opposite polarity.

Such differences evidence that query contribution is sensible to the application context. Indeed, the model captures the differences on the underlying datasets. For example, cubes of the Open workload being very large, many contributory queries are quite aggregated and have many filters for focusing in a specific portion of a cube. This explains the substantial weights of NoF, NoL and NoM. On the other hand, Enterprise data sources being simpler, both in the number of dimensions and levels, users tend to analyze the entire dataset (less filters), but at specific data granularity as required in their information needs. In this context, NoL and ADepth had more substantial weights than NoF and FDepth.

The models also capture labeling differences. For example, on the Open workload, the substantial weight of RI (relative identity) and in general of almost all relative features, reflect the expert's taste of contributory queries being similar to previous ones. On the contrary, on the Enterprise workload, the very high weight of QoI (quantity of information) as well as negative weights for relative features, translate expert's opinion of contributory queries providing new information instead of repeated one.

A larger interpretation of each model can be found in [Djedaini et al., 2019].

Models quality Table 2.7 reports models quality in terms of accuracy, precision and recall. The top part corresponds to the two contribution models described above, while the bottom part reports results of the cross-evaluation test.

Learning	Testing	Accuracy	Precision	Recall
Open	Open	0.880	0.960	0.902
Enterprise	Enterprise	0.800	0.817	0.831
Open	Enterprise	0.408	0.345	0.064
Enterprise	Open	0.860	0.862	0.998

Table 2.7: Quality of learned models for the Open and Enterprise workloads (top) and cross-evaluation (bottom)

As a conclusion of these experiments, our learning approach allows the learning of a definition of query contribution (difficult to be verbalized) that captures the characteristics of the underlying dataset and respects experts' judgment. However, the cross-evaluation experiment shows that our approach is not able, in this context and because of the inherent limits presented before, to learn a single contribution model that can be directly applied on all workloads.

Some additional results on artificial explorations of the Artificial workload are reported in [Djedaini et al., 2017b, Djedaini et al., 2019].

Models vs users' skills In order to confront our models to users' skills, we first compute average focus and average contribution for explorations of the Open workload, and then score explorations using KT model.

Average focus and contribution. For explorations labeled A, B and C, we obtain resp., an average focus of 0,241, -0,240 and -1,767, and an average contribution of 0,172, 0,028 and -0,548. Other statistics and complementary experiments (not reported here) confirm this tendency (see [Djedaini et al., 2019] for details).

We conclude that users who acquired knowledge (class A) conducted more focused explorations in average compared to the others. This reasoning is inversely true for explorations in class C. Class B is an intermediate situation, quite ambiguous, where it cannot be stated clearly that the skill has been mastered or not.

KT parameters estimation. In order to learn the KT model, we start by learning KT parameters, as explained in Subsection 2.3.2. It can be seen from Table 2.8 that the initial probability of writing a contributory query $P(L_0)$ is very low on the Open workload, which can be directly related to the fact that explorations have been performed by master students who knew very little about the data beforehand. Interestingly, and expectedly, $P(L_0)$ is higher for analysts of the Enterprise workload.

However, in both cases users have a sound theoretical background on OLAP exploration which in turn explains the relatively good probability, $P(T)$, to acquire the skill at each step of the exploration. Finally, the exploratory nature of OLAP analysis, combined with limited knowledge of the dataset, translates in the explorations by a lot of trials and errors that increased significantly the average probability and standard deviation of $P(G)$ and $P(S)$.

Parameter	Open	Enterprise
$P(L_0)$	0.085	0.238
$P(T)$	0.243	0.360
mean($P(G)$)	0.320	0.331
variation($P(G)$)	0.307	0.297
mean($P(S)$)	0.323	0.330
variation($P(S)$)	0.273	0.278

Table 2.8: Main parameters of continuous KT as learned on the Open and Enterprise workloads.

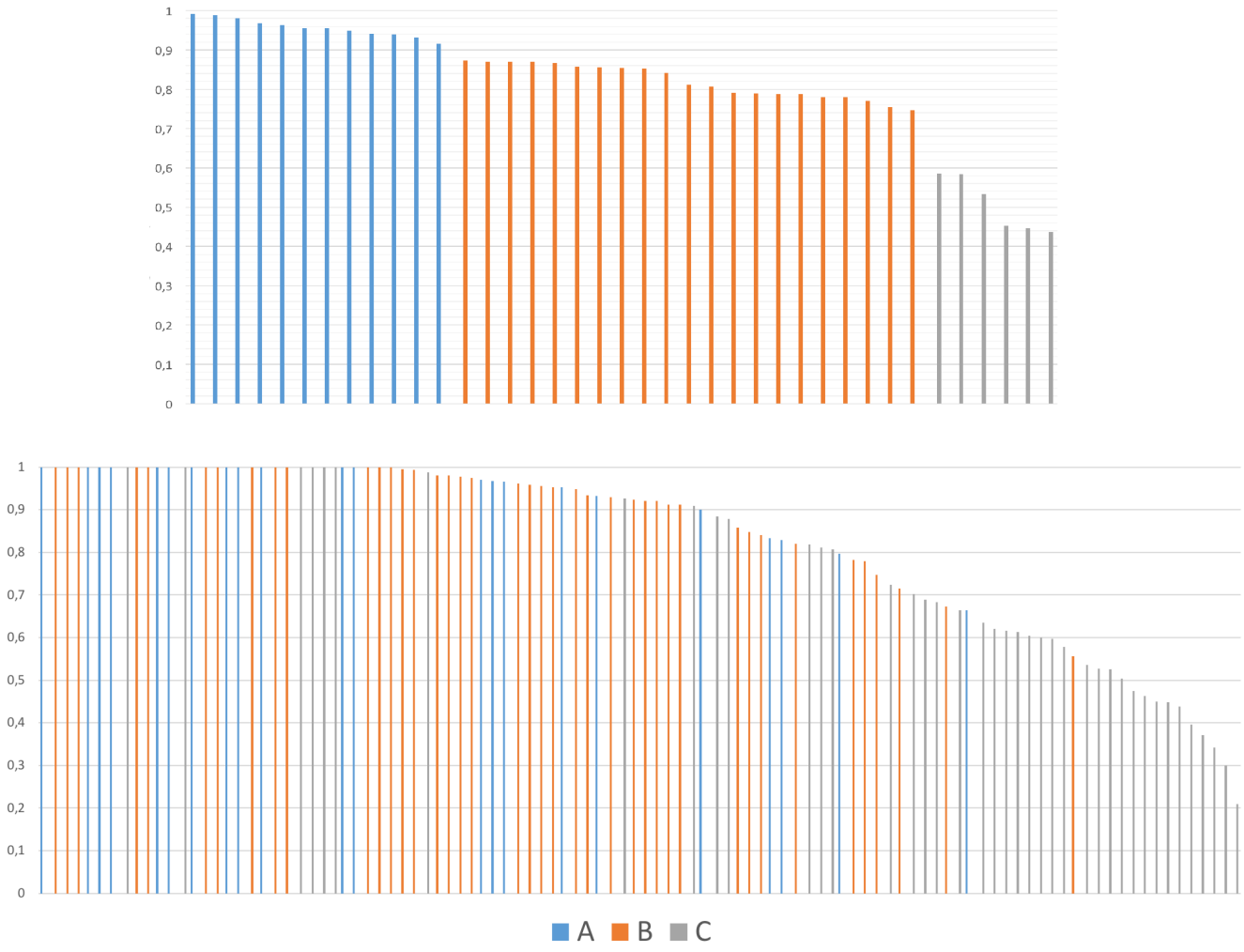


Figure 2.1: KT prediction of analysts' skills (scores) vs. experts evaluation (colors) for the Open (top) and Enterprise (bottom) workloads. Each bar corresponds to an exploration.

Skill prediction. Figure 2.1 represents explorations on the horizontal axis and the prediction of users' level of expertise provided by our KT model on the vertical axis. The colors represent the assessments made by experts.

For the Open workload, it appears clearly that our model provides consistent evaluations, giving users a rating that corresponds to the assessment made by the expert. The distinction between competent analysts (A and B) and non-competent analysts (C) is clearly marked, in contrast to the distinction between A and B which is less pronounced. This can be explained by the fact that it is difficult for an expert to distinguish between a good and a very good analyst. However, the distinction between a good and an unqualified analyst is intuitively much easier.

The result is more nuanced on the Enterprise workload, even though we retrieve a quite good distinction between competent and non-competent analysts (C-labeled explorations tend to cluster on the right of the chart, which corresponds to low scores).

Interestingly, if only one third of the explorations obtain a score greater than 0.9 for the Open workload, it is slightly more than 50% for the Enterprise workload, which reflects the average expertise of users. In other words, the KT model is able to correctly retrieve that Enterprise users are rather skilled compared to students.

Prediction quality. Finally, we evaluate our continuous KT based on a traditional RMSE score. RMSE has been shown to be the strongest performance indicator for binary KT with significantly higher correlation than Log Likelihood and Area Under Curve [Pelánek, 2015]. In our case, this RMSE score is computed as the difference between the expected contribution of each query in an exploration and the value predicted by the continuous KT for each of these queries. Our KT model obtains a RMSE score of 0.291 with a standard deviation of 0.181 for the Open workload, and of 0.238 with a standard deviation of 0.270 for the Enterprise workload.

Consistently with the literature on KT, we consider that these scores are rather good and we conclude that our model is effective at assessing if the skill *writing a contributory query* is acquired.

2.3.4 Discussion

This section proposed an approach to automatically assess the quality of OLAP queries and explorations.

Our approach for qualifying queries is based on a model of query contribution, built using supervised learning, which exploits a large set of query features. This model relates to the user's skill of writing queries that contribute to the exploration. Our approach for qualifying explorations is based on a model of skill acquisition that estimates the probability that a skill is mastered from a collection of opportunities to use the skill (i.e. the queries). We remark that we have used a similar principle to score sequences of book reviews [Megasari et al., 2018].

To our knowledge, our contribution is a pioneer of its kind. We successfully built a model, trained on a relatively large set of real explorations. We validated experimentally our model on a test set of real explorations. On top of that, we checked the coherence of our model by using it to detect how skilled is a data analyst.

We showed that automatic assessment of OLAP explorations is feasible and is consistent with the user's and expert's viewpoints.

Besides the experiments that validate the robustness of our models, we evaluated feature computation time. In average, the computation of all features for a given query is 695 milliseconds, which is negligible given that the average consideration time for query answers is $11,200$ milliseconds. These scores validate that is feasible to include focus and contribution computation in EDA support tools.

Many practical benefits of the proposed assessment technique can be envisioned. As it puts the user and their skills in the center of the data analysis activity, it can be seen as an important driver in the design of systems supporting EDA, as well as the corner stone of the development of benchmarks for such systems. In this direction, we proposed a framework for benchmarking exploratory OLAP support systems [Djedaini et al., 2016]. We showed that such a benchmark can be implemented using state-of-the-art techniques for data and user traces generation, and for metrics definition. We have validated the benchmark by proving that it correctly ranked a set of exploration strategies for which the behavior is well known.

In next studies (described in the following sections), we aim to relax the assumption of multidimensional schema and query language, and target SQL explorations over less normalized databases. In addition, a challenging direction is to switch to unsupervised learning, to avoid the need for manual labeling and the strong dependency on human expert annotation variability.

2.4 Segmentation of query workloads

Foreword

This section summarizes the master thesis of Willeme Verdeau, that I supervised. It also concerns the master projects of Yann Raymond and Aboubakar Sidikhy Diakhaby, co-supervised with Patrick Marcel.
The proposal was published at DOLAP [Peralta et al., 2019b] and extended at Information Systems [Peralta et al., 2020].

In this section, we present an approach for segmenting a query workload into explorations. Our work aims at finding the best way of segmenting a query workload, upon which little is known (no timestamps, no ground truth, no database instance), into meaningful, coherent explorations.

We investigate three alternatives for session segmentation:

- unsupervised learning: our first method is based only on similarity between contiguous queries,
- supervised learning: our second method uses transfer learning to reuse a model trained over a workload where ground truth is available,
- weak supervision: our third method uses weak labelling to predict the most probable segmentation from heuristics meant to label a training set.

Considered workloads. We experiment with the SQLShare workload, which, as reported in [Jain et al., 2016], is the only one containing primarily ad-hoc hand-written queries over user-uploaded datasets. A preliminary session segmentation (contiguous queries of a given user) resulted in some extremely long sessions (maximum of 937 queries) with 26% of queries having nothing in common with their immediate predecessor. Then, session segmentation appears as an unavoidable step for any explorative usage of SQLShare.

In addition to SQLShare, we experiment with the Open and Enterprise workloads, and with their concatenation, Concatenate workload. These workloads, while containing a particular case of queries (star-join queries), are interesting because a ground truth (the set of queries corresponding to each exploration) is available, allowing the evaluation of our approach.

Table 2.9 (top part) provides an overview of the Workloads in terms of number of queries, sessions and explorations (when available). We remark that the Open workload contains long[11] sessions concerning few explorations while the Enterprise workload contains shorter sessions concerning more explorations. In addition, note that in terms of queries per session, the SQLShare workload is similar to the Enterprise one.

Query features. We represent a query as a vector of query features (subset of the intrinsic and relative features presented in Section 2.2). We focus on features counting frequently-used query fragments (namely, number of projections (P), selections (S), aggregations (A) and tables (T)) and features capturing common fragments (namely, number of common projections (NCP), selections (NCS), aggregations (NCA) and tables (NCT)). Two additional comparison features are experimented: edit distance (RED) and Jaccard index (JI) as they combine several query fragments.

[11] Sessions length is actually dependent on the GUI used; while third party OLAP tools, like Saiku, log a new query for each user action (including intermediate drag-and-drops), the SAP prototype only logs final queries.

	Open	Enterprise	SQLShare
Nb of sessions	16	24	451
Nb of explorations	28	104	
Nb of queries	941	525	10,668
Avg queries per session	58	21	24
Avg queries per explor.	34	5	
Avg explor. per session	2	4	
Avg and range of P	3.62 [1,7]	2.18 [0,6]	9.14 [1,509]
Avg and range of S	3.61 [0,26]	1.79 [0,5]	1.19 [0,83]
Avg and range of A	1.34 [1,4]	1.14 [0,5]	0.39 [0,48]
Avg and range of T	3.28 [1,7]	2.03 [1,4]	1.50 [0,84]
Avg and range of NCP	3.16 [0,7]	1.34 [0,4]	4.92 [0,509]
Avg and range of NCS	3.12 [0,25]	1.03 [0,5]	0.59 [0,82]
Avg and range of NCA	1.17 [0,4]	0.77 [0,3]	0.20 [0,48]
Avg and range of NCT	2.97 [0,7]	1.46 [0,4]	0.85 [0,83]
Avg and range of RED	3.85 [0,19]	2.09 [0,25]	10.82 [0,1020]
Avg and range of JI	0.57 [0,1]	0.79 [0,1]	0.45 [0,1]

Table 2.9: Length and features of Open, Enterprise and SQLShare workloads

Table 2.9 also summarizes feature extraction. We remark that queries in the Open and Enterprise workloads concern a quite small number of projections, selections, aggregations, and tables. Conversely, SQLShare queries, in average, are richer in terms of projections (with high variations among queries)[12], but contains less aggregations, selections and tables. Regarding relative features, except for the number of common projections, most features show that queries are less similar than in the other workloads. Relative edit distance (RED) and Jaccard index (JI) illustrate that queries are more similar in the Enterprise workload.

Next subsections describe the three proposed segmentation methods.

2.4.1 Similarity-based session segmentation

Intuitively, our idea is to compare contiguous queries in a session and segment when queries are dissimilar enough. Based on the query features previously described, we investigate 5 similarity indexes:

Edit Index. It is based on the Relative Edit Distance (RED) query feature. For normalizing, RED is translated to the [0,1] interval, considering similarity is 0 after a given number of operations (arbitrarily set to 10).

$$EditIndex(q_k, q_{k-1}) = max\{0, 1 - \frac{RED(q_k, q_{k-1})}{10}\} \tag{2.1}$$

Jaccard Index. It corresponds to Jaccard Index (JI) feature, which is normalized by definition.

$$JaccardIndex(q_k, q_{k-1}) = JI(q_k, q_{k-1}) \tag{2.2}$$

Cosine Index. It is calculated as the Cosine of vectors consisting of 8 query features, namely, P, S, A, T, NCP, NCS, NCA, and NCT. Let $x = \langle x_1, \ldots, x_8 \rangle$ and $y = \langle y_1, \ldots, y_8 \rangle$ be the vectors for queries q_k and q_{k-1} respectively.

[12]The use of * wildcard has great influence in such variations.

$$CosIndex(q_k, q_{k-1}) = \frac{\sum x_i.y_i}{\sqrt{\sum x_i^2 . \sum y_i^2}} \tag{2.3}$$

Common Fragments Index. It is calculated as the number of common fragments normalized to the [0,1] interval and considering similarity is 1 when there are more than 10 common fragments (arbitrarily set).

$$CFIndex(q_k, q_{k-1}) = min\{1, \frac{NCF}{10}\} \tag{2.4}$$

where $NCF = NCP(q_k, q_{k-1}) + NCS(q_k, q_{k-1}) + NCA(q_k, q_{k-1}) + NCT(q_k, q_{k-1})$.

Common Tables Index. It is calculated as the number of common tables. We wanted this index to be relative to the user's session ; this is why normalization here is specifically achieved in relative terms, by dividing by the highest number of tables in the session.

$$CTIndex(q_k, q_{k-1}) = \frac{NCT(q_k, q_{k-1})}{max\{T(q)|q \in session(q_k)\}} \tag{2.5}$$

Note that these indexes calculate complementary aspects of query similarity and are normalized in different ways. Our intention is to capture different points of view and therefore to deal with different situations. Edit Index and Common Fragment Index count differences (resp., common fragments) as absolute values (resp. normalized with a given threshold). Jaccard Index is a compromise of the previous ones, computing the ratio of common fragments. Cosine Index is computed using features values instead of comparing sets of fragments; it captures the variability in query complexity. And finally, Common Table Index responds to the intuition that common tables have more impact than the other common fragments, and it is normalized with respect to the number of tables used in the user's session.

Example 2.2 *Figure 2.2 depicts the similarity indexes for 3 sessions of the SQLShare workload, having different sizes. Looking at Session 28, the shorter one, it seems quite clear that the session may be split in two parts, by cutting between queries 4 and 5. All similarity indexes agreed. Things are less evident for Session 0. One split seems evident (at query 31), but some others may be discussed (e.g. at queries 29 and 12). Decision to split the session will depend on what similarity thresholds to use for the indexes. Finally, Session 18 presents a first part, with a focused analysis, via similar queries, and a second part, more exploratory, with varied queries. Even if indexes do not always agree, their majority seems to indicate a tendency.* □

In practice, our approach can be summarized as follows: For each pair of consecutive queries: (i) compute query similarity according to the proposed similarity indexes, (ii) compare the obtained similarity values with their respective thresholds, obtaining a set of votes for "CONTINUE" (do not segment) or "SEGMENT" (segment and start a new exploration). The decision (to keep consecutive queries together, or to segment) is taken by majority.

Similarity thresholds are experimentally tuned on the distribution of values of each similarity index and experiments on workloads where there is a ground truth (see details in [Peralta et al., 2020]).

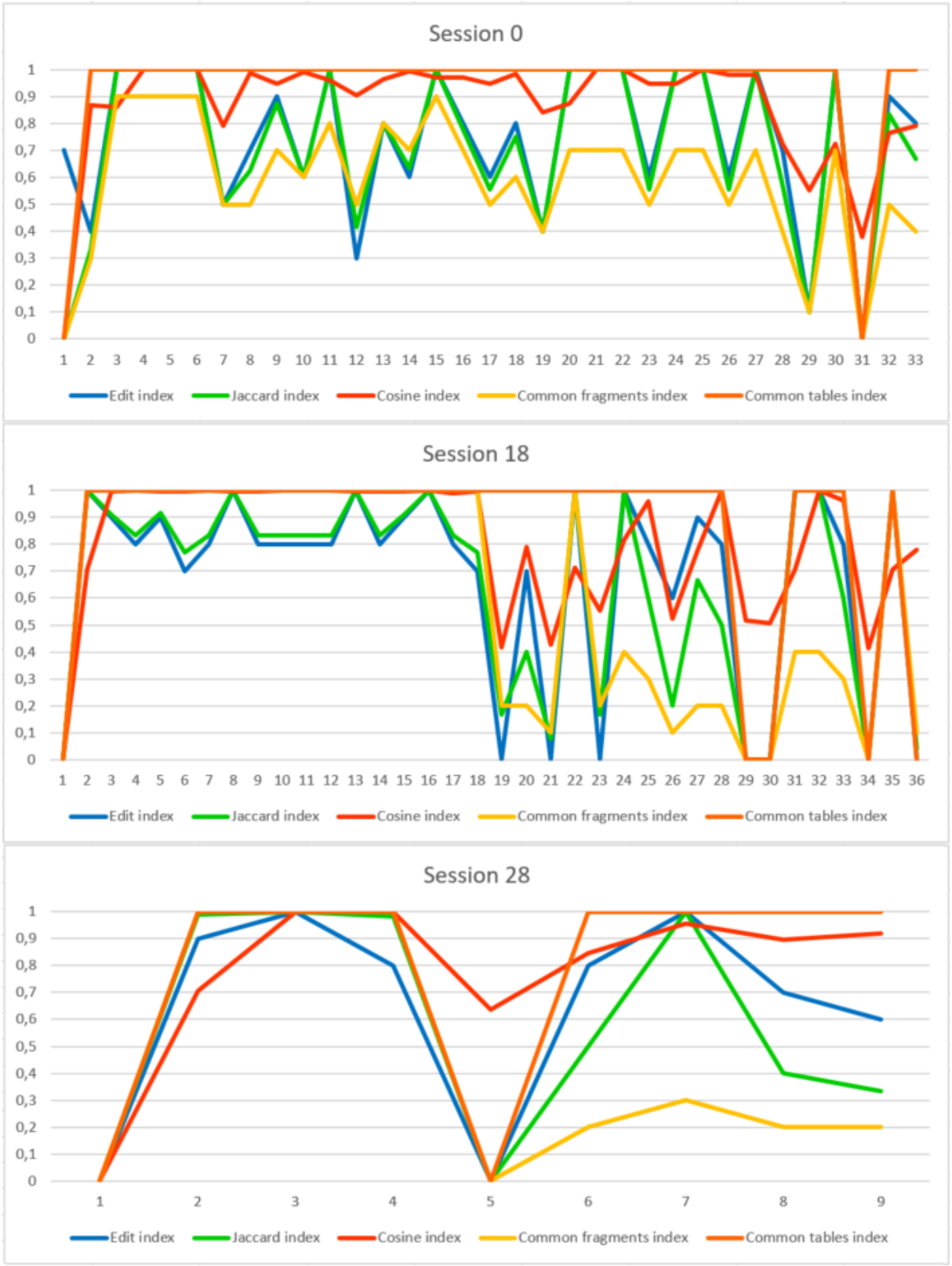


Figure 2.2: Comparison of similarity indexes for 3 sessions.

2.4.2 Transfer learning based session segmentation

Our second method for segmenting the SQLShare workload is based on transfer learning, that consists of using supervised learning to tune a model over a labelled dataset and use this model over a dataset for which no ground truth is available. We first introduce the basics of transfer learning, and then describe our approach.

Transfer learning. Classical supervised machine learning supposes large collections of previously collected labeled training data, to build effective predictive models. When labeled data is scarce, semi-supervised approaches may be used to build classifiers over a large amount of unlabeled data and a small amount of labeled data. Still, such approaches assume that the distributions of the labeled and unlabeled data are the same. Transfer learning, however, aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task, while allowing the domains, tasks, and distributions used in training and testing to be different. Transfer learning situations differ in what, how and when to transfer [Pan and Yang, 2010].

In our context, having no ground truth for the SQLShare workload, but having ground truth for other workloads, and considering the difference in feature correlation between SQLShare and the other workloads (not reported here), allows to model session segmentation as a classification task, and use transfer learning. Precisely, we will consider learning a classifier over ground truth workloads as a source task, and learning a classifier over SQLShare as the target task. According to the typology introduced in [Pan and Yang, 2010], this is a case of transductive transfer learning setting, where the source and target tasks are the same, while the source and target domains are different, but the feature spaces between domains are the same. In that case, learning a model that can generalize to the target workload demands to remove the sample selection bias due to the fact that source data and target data are drawn from different distributions. This can be achieved by reweighting the source data after having estimated the probability of appearance of each sample of the source workload in both the source and the target workload, which can be done for instance with density ratio estimation [Sugiyama et al., 2007, Huang et al., 2006].

Binary classification with linear SVM. We formalize the problem of workload segmentation as a supervised classification task, in the spirit of what we did for learning query quality (described in Section 2.3) and reusing the workloads with ground truth. We represent a query by a set of features that are the most correlated to the ground truth. Our objective is to learn a linear combination of the features that separates queries starting an exploration, from those continuing an exploration. To this end, each vectorized query of the ground truth is associated with a binary label: SEGMENT (a segmentation to be found), and CONTINUE (no segmentation).

To learn our model, we trained a binary classifier over the workload, removing sample selection bias by reweighting samples using kernel-mean matching (KMM) [Huang et al., 2006]. We chose a linear SVM classifier since this proved effective in previous section. The model is learnt using 10-fold cross validation, choosing its best hyperparameter via randomized search.

Since we believe that the workloads are likely to be heavily unbalanced towards the CONTINUE label, we tested various balancing strategies while training the model, aiming at improving classification accuracy. We compared several methods on the basis of their respective accuracy and F1-measure, over a 10-fold cross-validation: either random undersampling of majority class, or oversampling of minority class. In the last case, several heuristics have been tested: random oversampling, 3 variants of SMOTE (with different approaches to sample borderline points between classes) or ADASYN [Batista et al., 2004].

Once the best hyperparameter is obtained, the model is eventually trained over the full reweighted Concatenate workload, to be applied over the target SQLShare workload.

2.4.3 Weak labelling and generative model

Instead of directly learning a transferable model from a labelled workload, our third approach uses a generative model to predict the labels of the unlabelled workload. To this end, we resort to weak supervision, a labeling technique consisting of using noisier or heuristic sources of labels to avoid hand-labeling data.

We use Snorkel [Ratner et al., 2017], a weak supervision system that (1) lets users write labeling functions (LFs), (2) applies the LFs over unlabeled data and learns a generative model to combine the LFs' outputs into probabilistic labels, and eventually (3) allows to use these labels to train a discriminative classification model.

Snorkel is intended to work over unstructured data. Labeling functions take as input a candidate object, representing a data point to be classified. Each candidate is a tuple of context objects, which are part of a hierarchy representing the local context of the candidate [Ratner et al., 2017]. Typically, a candidate is a pair of named entities and the context is a sentence in which they both appear, this sentence itself being part of a document, the set of documents being the dataset to be labelled.

We adapt to Snorkel's data model by considering each session of the labelled workload as a context, and each pair of consecutive queries in a session as a candidate.

We write simple (potentially contradictory) labelling functions using the features and indexes extracted from the workloads. To maximize agreement between labelling functions, we grouped them and select the best subset of each group in terms of F1-measure, when trained over the labelled workload. We then merge the best subgroups and repeat this process until the score no longer improves. We give below a brief description of our labelling functions.

Labeling functions. We implemented 21 labelling functions, each one using one of the relative features or indexes extracted from the workload. Considering that query workloads can be very different, our objective was to define functions that capture, through simple heuristics, intuitive properties of pairs of queries, and to remain independent from the workload. As with the previous approach, we use a binary labelling scheme (CONTINUE, SEGMENT).

Our first group of functions consists of one function per index (edit index, etc.), all being based on the same algorithm: if the index is greater than 0 then the pair is assigned label CONTINUE, otherwise label SEGMENT is assigned.

Our second group of functions implements a precision and a recall indicator for each of the 4 relative metrics (NCP, NCS, NCA, NCT), resulting in 8 Functions. For such a relative metric, say NCP, recall (resp. precision) is computed as $\frac{NCP}{NP_f}$ (resp. $\frac{NCP}{NP_s}$) where NP_f (resp. NP_s) is that of the first (resp. second) query of the pair. All labelling functions are then based on the same algorithm: if recall (resp. precision) equals 1 then the pair is assigned label CONTINUE, else if it equals 0, then label is SEGMENT. Otherwise the function does not assign any label.

Our third and last group is a second implementation of precision and recall for all 4 relative metrics (another 8 functions), favoring the attribution of the CONTINUE label, as follows: if recall (resp. precision) is not 0 then label is CONTINUE, otherwise it is SEGMENT.

2.4.4 Experiments and results

In this section we report the major findings of our experiments for testing the three proposed methods, referred hereafter as Voting, Transfer and Weak-labelling, resp. We test our methods on the Open, Enterprise, Concatenate and SQLShare workloads and also report the agreement between them.

Protocol and baseline. The SQLShare, Open and Enterprise workloads are preprocessed for extracting query fragments and computing query features (as described in Section 2.2). Similarity indexes are computed as described in Subsection 2.4.1.

The input of our segmentation methods is a CSV file per workload (SQLShare, Open, Enterprise and Concatenate), each line describing a query by means of: query id, session id, query features, similarity indexes, and ground truth when available (labels SEGMENT and CONTINUE). The output of each method is an additional column in each file, indicating the segmentation (labels SEGMENT and CONTINUE).

In experiments with ground truth, both columns (ground truth and segmentation) are compared in order to evaluate the effectiveness of each method. We compute four classical quality metrics, defined as follows:

- *Accuracy* measures the ratio of queries having the same label.
- *Precision* measures the ratio of queries coinciding in SEGMENT label among the queries labeled SEGMENT in the obtained segmentation.
- *Recall* measures the ratio of queries coinciding in SEGMENT label among the ones having SEGMENT label in the ground truth.
- *F-measure* computes the harmonic average of precision and recall.

Our baseline is a naive method always predicting the majority class (i.e., always predicting CONTINUE and never predicting SEGMENT). It obtains good values for accuracy (97% for Open, 82% for Enterprise and 91% for Concatenate). However, such baseline obtains 0 as score for F-measure (since there is no SEGMENT prediction). We then simply use prediction of the majority class as a baseline for accuracy.

In order to compare our approach to the one used in the literature, we implement an additional method that segments users' sessions when there is a 30-minutes delay between queries.

Implementation and setting. Methods are implemented in Python; code and data are available from Github[13]. We tune several parameters and heuristics for each method, keeping the best configuration, based on knee detection. For Voting method, we tune and compare several thresholds for similarity metrics. For Transfer method, we select the features the most correlated with the ground truth in order to reduce dimensionality, and experiment several balancing techniques. For Weak-labelling method, the first task consist in the selection of the most appropriate subset of labelling functions. We select the best subset, in the sense of F-measure, over the Concatenate workload.

Segmentation quality. We first test our methods for the workloads with ground truth. For each workload, we compare the obtained segmentation to the ground truth, measuring segmentation quality in terms of accuracy, precision, recall and F-measure.

Voting method is tested on the three workloads; results are reported in Table 2.10. Conversely, Transfer and Weak-labelling methods are only tested on the Concatenate workload in order to have a larger training set. A comparison of the results of the three methods, and the baseline, on the Concatenate workload, is reported in Table 2.11.

As expected, results are very good in terms of accuracy, mainly explained because classes are unbalanced and quite good in terms of F-measure. We note that of all three methods, Voting, with its simple underlying idea based on similarity indexes and thresholds, outperforms the other ones, while being unsupervised. Interestingly, a correlation study shows that Jaccard index is

[13] `https://github.com/patrickmarcel/SQLWL-segmentation`

	Voting			Timestamp
	Open	**Enterprise**	**Concatenate**	**Open**
Accuracy	0.99	0.92	0.97	0.99
Precision	1	0.75	0.79	1
Recall	0.75	0.81	0.80	0.64
F-measure	0.86	0.78	0.80	0.78

Table 2.10: Segmentation results for the Voting method on the 3 workloads and for the Timestamp-based method (rightmost column)

	Voting	**Transfer**	**Weak-labelling**	**Baseline**
Accuracy	0.97	0.959	0.959	0.91
Precision	0.79	0.764	0.76	
Recall	0.80	0.758	0.766	
F-measure	0.80	0.761	0.763	

Table 2.11: Segmentation results for the 3 proposed methods and the baseline

the most correlated to the ground truth for all workloads, and it is also the most correlated to the final vote, so being the most influencing index. In addition, all methods obtain much better accuracy than the simple baseline predicting the majority class (CONTINUE).

The Open workload, the only one containing timestamps, is also compared to the Timestamp-based approach used in the literature. Results are reported in the right-most column of Table 2.10. They are comparable in terms of accuracy and lower in F-measure. Note that 1 for precision means that all cuts found are also breaks in the ground truth. In other words, there are no big delays inside explorations, which makes sense. However, the timestamp-based approach fails to detect 36% of the breaks (when the user changes its topic of study in a briefer delay).

Results on SQLShare. The three methods, with their best configurations, are used for segmenting the SQLShare workload.

The Voting method split the initial 451 sessions in 3,075 explorations. In the absence of ground-truth, we present in Table 2.12 a comparison of some features before and after session segmentation using the Voting method. A first remark concerns session length: extremely large sessions (maximum of 937) are split (new maximum is 98 queries). Indeed, more than half of the sessions are not fragmented and at 3rd quartile 1 session is split in 3 explorations. Some long and anarchic sessions (such as the one counting 937 queries) are split in a multitude of explorations. We can also highlight an increase in the average number of common query fragments (NCP, NCS, NCA, NCT) per session. This increase is quite regular and visible for all quartiles. Relative edit distance (RED) and Jaccard Index (JI) also improve, as expected.

Trained over the Concatenate workload, using KMM re-weighting, and applied over the SQLShare workload, the Transfer method obtained 3,420 explorations. Analogously, the Weak-labelling method produced 3,175 explorations.

Importantly, all methods agree on finding more than 26% of segmenting, consistently with our preliminary analysis of the SQLShare workload (as there are 26% of queries having nothing in common with their immediate predecessor). In addition, in 93% of the cases methods have full agreement and in 98% of cases the Voting method agree with at least one of the other methods. In the remaining 2% of cases, Voting keeps queries together while the other methods propose to segment. Indeed, Voting method detects the less explorations, 3,075 against 3,174 and 3,420 for the other two methods.

	Avg	Stddev	Min	25pc	50pc	75pc	Max
Before segmentation							
Nb queries	23.65	75.05	1	2	4	13.50	937
Avg NCP	6.14	19.54	0	1	2.37	5.21	306
Avg NCS	0.44	0.74	0	0	0.17	0.71	7
Avg NCA	0.19	0.45	0	0	0	0.17	4
Avg NCT	0.93	0.60	0	0.67	0.97	1	4
Avg RED	8.49	17.33	0	2.71	5.35	8.55	205
Avg JI	0.52	0.27	0	0.33	0.53	0.69	1
After segmentation							
Nb queries	3.47	5.73	1	1	1	3	98
Avg NCP	6.98	17.21	0	1.33	3	7.50	509
Avg NCS	0.64	1.96	0	0	0	1	55
Avg NCA	0.29	1.72	0	0	0	0	48
Avg NCT	1.18	3.08	0	0.8	1	1	82
Avg RED	8.02	20.70	0	1.67	3.48	7	508
Avg JI	0.61	0.27	0	0.41	0.64	0.84	1

Table 2.12: Comparison of number of queries and average relative features per session, before and after segmentation

2.4.5 Discussion

This section addressed the problem of segmenting sequences of SQL queries into meaningful explorations when only the query text is available, and it is not possible to rely on timestamps.

We characterized queries as a set of simple features and defined five similarity indexes with respect to previous queries in the session. A simple unsupervised method, based on the similarity indexes with voting strategy, allowed to split long and heterogeneous sessions into smaller explorations where queries have more connections. This method tunes similarity thresholds based on knee detection and uses no labels nor expert knowledge. We investigated two additional methods, exploiting supervised and weak-supervised learning techniques. Experiments showed a strong agreement among the 3 methods; the best results, in terms of accuracy and F-measure over workloads with ground truth, being achieved by the simple unsupervised method.

From a practical point of view, the Voting method is also easier to implement as it does not require any training with labeled workloads nor labelling functions.

In next studies (described in the following section), our choice is to use the explorations found by the Voting method, as it is the simplest one, does not need any labelling and achieves good results. The high agreement with the other methods reinforces our decision.

Our approach can be easily extended with other query features and other similarity indexes, in particular for considering each query in the context of its session (not only comparing it to its immediate predecessor) and exploiting query answers. Further similarity indexes may be deduced from such features.

In this work, we have only considered SELECT statements from the SQLShare workload. However, there are 469 remaining statements that represent updates and inserts. They are interesting as they may represent intermediate or partial results. In addition, we notice that some statements are attempts to deal with formatting problems and data quality issues. Their parsing and inclusion may be an interesting extension of this work.

We hope that our segmentation approach could help improving a variety of novel log-based applications, from the measurement of the quality of SQL explorations, the detection of specific exploratory activities, the learning of users' analysis behavior, the discovery of latent users' intents, or the recommendation of forthcoming exploration queries.

2.5 Learning analysis patterns

Foreword

This section summarizes part of the PhD thesis of Clément Moreau, co-supervised with Thomas Devogele and Laurent Etienne. It also concerns the master projects of Clément Legroux and Mohamed Ali Hamrouni, that I supervised.
The proposal was published at DOLAP [Moreau et al., 2020c, Moreau and Peralta, 2021] and extended at Information Systems [Moreau et al., 2022].

In this section we present an approach for learning analysis patterns in a query workload containing explorations devised by real users.

We propose a similarity measure tailored for comparing explorations and pair it with a off-the-shelf clustering algorithm. We aim at obtaining a set of clusters of similar explorations (both in terms of operations and complexity), each cluster revealing a pattern of analysis behavior.

Considered workloads. We firstly consider four workloads of analytical queries, namely Artificial, Ipums, Open and Security. The former three contain multidimensional OLAP queries, represented via <group by, selection, measure> triplets. The latter contains simple SQL queries generated by an analytical tool, thus being close to OLAP queries. Actually, the query model is slightly richer, also providing projections and order by attributes.

We then generalise to regular SQL queries and experiment on the SQLShare workload, reusing the Voting segmentation method described in Section 2.4. We remark that in the SQLShare workload, length of explorations (i.e. the number of queries in an exploration) follows the Zipf's law. In particular, 1,379 explorations are one-shot, i.e. they contain only one query.

For the Artificial and Ipums workloads, we have some knowledge describing analysis style that can be used as a ground truth, namely, the templates used for the generation of the Artificial workload (Slice And Drill, Slice All, Exploratory, Goal Oriented) and preliminary labels indicating analysis style of the Ipums workload (Focus, Oscillate-Focus, Oscillate, Fix, Atypical); they are described in Appendix A. These templates and labels, even not being a real ground truth for our method, provide a guide for comparison.

Query features We compute four types of query vectors, two concerning operations and two concerning complexity.

- An operation vector is a 12-dimensional vector concerning operations between queries, namely, the number of added and deleted projections (+P, -P), selections (+S, -S), aggregations (+A, -A), tables (+T, -T), group by expressions (+G, -G) and order by expressions (+O, -O).
- A reduced version of the operation vector is used for analytical queries. It concerns operations on selections, aggregations (i.e. measures) and group by expressions (+S, -S, +A, -A, +G, -G) and two additional features indicating the level of aggregation and filtering (ADepth, FDepth).
- A length vector is a 5-dimensional vector indicating query length in terms of number of characters (C), attributes (B), tables (T), sub-queries (Q) and functions (F).
- A clause vector is a 4-dimensional vector indicating the types of complex clauses used in a query. It counts the number of advanced joins (J), set operators (U), advanced clauses (V) and expert clauses (E).

Aggregated operation and clause vectors are also computed.

2.5.1 Query and exploration similarity

We use cosine similarity for computing similarity between query vectors. This measure is well suited to compute the similarity between two vectors and is normalized in $[0, 1]$. In this way, it favors more the nature of SQL operations and clauses than their number. To deal with zero vectors, which are frequent in the SQLShare workload, we set border cases as follows: (i) zero vectors are considered identical (similarity is 1), and (ii) one zero vector is considered as completely different from a non-zero vector (similarity is 0). Formally, given two vectors v and v', cosine similarity is calculated as follows:

$$\cos(v, v') = \begin{cases} 1 & \text{if } \|v\| = 0 \text{ and } \|v'\| = 0 \\ 0 & \text{if } \|v\| = 0 \text{ xor } \|v'\| = 0 \\ \frac{v \cdot v'}{\|v\|\|v'\|} & \text{else} \end{cases} \tag{2.6}$$

In order to compare explorations, we propose a Contextual Edit Distance (CED) tailored to the comparison of semantic sequences.

CED is a generalization of the Edit Distance, adapting cost computation to typical characteristics of semantic sequences[14]. In particular, CED answers the following requirements:

1. *Context-dependent cost*: Edition cost depends on the similarity of nearby elements. The more similar and closer the elements, the lower the cost of operations,
2. *Repetition*: Edition of repeated close elements has low cost.
3. *Permutation*: Similar and close elements can be exchanged with a low cost.

Example 2.3 *Consider an exploration reflecting an exploratory behavior at the beginning (many changes in measures and group by set) and more focus at the end (drilling and filtering). We can sketch it as follows (where G, S and A means group-by levels, selections and aggregations (measures), + means addition and - means deletion; we skip other query features for simplicity):* $\langle$ *+G+A, +A, +A, +G, +A-A, -A+G, -G+G, +S+G, +S+G, +S, +S* $\rangle$.

Consider the insertion of a query adding an additional measure (+A). The edition cost should be low if the query is inserted at the beginning (as it is similar to near queries), even lower at positions 2 to 4 (because repeating the same operations), but high at the end. □

This requirements ensure that explorations reflecting a given pattern (e.g. sequences of drill-downs) are judged to be very similar no matter the exploration length (i.e. how many drill-downs) nor the underlying data (which data was drilled-down).

We describe CED computation as defined in [Moreau et al., 2020b] and tuned in [Moreau et al., 2020c]. Firstly, CED modifies the cost function γ of Edit Distance to take into account the local context of each element in the sequence. Consider contextual edit operations of the form $O = (o, e, q, k)$, denoting the operation $o \in \{\texttt{add}, \texttt{modify}, \texttt{delele}\}$ on exploration $e = \langle q_1, ...q_n \rangle$ at index k by query q. Let $\mathcal{O}$ be the set of all possible contextual edit operations, the cost function $\gamma : \mathcal{O} \rightarrow [0, 1]$ is defined as:

$$\gamma(O) = 1 - \max_{i \in [1,n]} \{sim(q_i, q) \times v_i(O)\} \tag{2.7}$$

where: sim is the similarity measure between two queries and $v(O) \in [0, 1]^n$ is a contextual vector which quantifies the notion of proximity between queries. Usually, bigger $|i - k|$ is, smaller $v_i(O)$. As in [Moreau et al., 2020c], we use:

[14] In our case, sequences are explorations and their elements are queries.

$$v_i(O) = \exp\left(-\frac{1}{2}\left(\frac{2\sqrt{k+1}(i-k)}{|e|}\right)^2\right)$$

CED is computed as Edit Distance, using dynamic programming and Wagner-Fisher algorithm [Wagner and Fischer, 1974].

2.5.2 Indicators for clustering analysis

Other research communities, in particular mobility science [Miller, 2017, Parent et al., 2013, Kontarinis et al., 2021], study human behavior represented as sequences of actions. Data exploration can be viewed through the prism of mobility science [Hägerstraand, 1970]. Indeed, an exploration is a sequence of user's queries, where the movement is no longer conducted in space but in the *data space*.

Thus, many indicators proposed for the analysis of mobility sequences can be reused or adapted for the study of sequences of queries. Mobility researchers explored sequences of activities and tested the existence of simple universal rules underlying human movement like travel distance, top ranked visited locations, predictability of human activity and origin-destination flows, mainly studying recurring patterns/regularity in the sequence or clustering mobility behavior ([Barbosa et al., 2018] presents an important survey). In substance, results show that mobility is strongly characterized by exponential distribution (e.g. heavy-tailed, Zipf) and that people constantly exploit a small set of repeatedly visited locations.

This capacity to explain models, both for practical and ethical issues, is a crucial point for the understanding of machine learning models. With this aim in mind, Guidotti et al. [Guidotti et al., 2019] suggested some techniques, partially borrowed from these above, like statistical methods and prototype selection elements, to explain black box systems in order to make their results more interpretable and understandable. In line with the vision of these techniques, we believe that the elaboration of indicators is essential to understand and explain discovered behavior in complex clusters.

Inspired by these considerations, we propose to adapt a set of indicators from mobility mining to analyse data explorations. These complementary techniques, summarized in Table 2.13, highlight different aspects of explorations.

Techniques	**Description**	**Visual. method**
Statistical distribution		
Length distribution	Frequency distribution of sequence length	Boxplot
State distribution	Frequency distribution of elements inside the sequences	Barplot
Vector description		
ℓ_1 norm	Sum of vector coordinates. $\|v\|_1 = \sum_{i=1}^{n} v_i$	Boxplot
Correlation	Correlation of vector dimensions	Correlogram
Component analysis	Frequency of vector components	Barplot/Stackplot
Transitions		
Origin-Destination matrix	Number of transitions from vector q_i to q_j	Chord diagram
Scattering and outliers		
UMAP	Dimensional reduction. Visualization of complex elements in 2D Euclidean spaces with a preservation of local topology	Euclidean projection

Table 2.13: Indicators for sequence and vector analysis and typical visualization methods

2.5.3 Experiments and results

In this section we describe our experiments for clustering explorations and report the major findings about analysis patterns and users' skills.

In what follows, we consider four workloads of OLAP queries: Open, Security, Ipums and Artificial (the latter two having a ground truth), and a workload of SQL queries: SQLShare. As these workloads concern users with varied analytical skills and using varied query tools, we aim at discovering different types of patterns.

We pair CED to an off-the-shell clustering algorithm and we test it against these workloads.

Protocol. Query workloads are preprocessed for segmenting sessions (as described in Section 2.4), and extracting query fragments and computing query features (as described in Section 2.2).

In order to cluster explorations in each workload, we execute an off-the-shelf clustering algorithm using CED as distance function. For comparison, in workloads with ground truth, we execute the same clustering algorithm with two alternative distances: (i) the classical Edit Distance (henceforth dubed ED) as a baseline, and (ii) Aligon et al.'s distance [Aligon et al., 2014b] (henceforth dubed AD), a state of the art metric for session similarity.

We perform several experiments, using different subsets of query features and consequently, obtained several sets of clusters. Concretely, we investigate 3 clustering variants:

1. clustering of OLAP explorations using only operation vectors,
2. clustering of SQL explorations using only operation vectors, and then analysis of complexity per cluster, and
3. clustering of SQL explorations using all available features (i.e. those of both operation and complexity vectors).

The first variant aim to study analysis behavior captured by OLAP operations, in a controlled configuration (OLAP). We report several clustering quality scores[15]: Firstly, when we have a ground truth, we compare clusters found with CED, to those obtained with ED and AD, reporting Adjusted Rand Index (ARI) and V-measure (harmonic mean of clusters homogeneity and completeness) scores. In addition, for the four workloads, we report intrinsic cluster quality scores, aiming to balance the number of clusters, cluster diameter (the distance between the farthest objects in the cluster) and mean Silhouette Coefficient[16]. Indeed, too few clusters will mix different behaviors, too many clusters will overfit users' behavior.

The remaining variants study SQL explorations, also considering query complexity. The second variant aims to investigate if SQL operations determine query complexity, and in a general way, looking if there is a relation between both types of features. The third variant investigates the mixing of features coming from operation and complexity vectors in a unique clustering. In both cases, given the large number of one-shot explorations (i.e. containing a unique query) in the SQLShare workload, we separately cluster one-shot explorations and longer ones. This separation aims to further analyse longer explorations, revealing richer patterns.

Finally, we use the indicators described in Subsection 2.5.2 for analysing the obtained clusters. In some cases, we also study the medoids of each cluster (the exploration that is the most similar to all other explorations in the cluster) for providing an explanation of median behavior.

[15]Metrics for clustering performance evaluation are well described in [Pedregosa et al., 2011] and `https://scikit-learn.org/stable/modules/clustering.html` sect. 2.3.10.

[16]Silhouette Coefficient is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Silhouette scores are merely informative in out tests, as the metric is more adapted to hyper-spherical clusters.

Implementation and setting. Our methods are implemented in Python, using SCIPY, SKLEARN and MATPLOTLIB libraries. Code and data for experiments on OLAP workloads are available from Github[17]; those on SQLShare are available from our Python notebook[18].

In our first experiments on OLAP workloads, we use a hierarchical clustering algorithm, which provides more flexibility than hyper-spherical and density-based algorithms when one do not know, a priori, the form of clusters, nor their density. In addition, it outputs a dendrogram that allows to parameter the setting of number of clusters and eases the visual analysis of clusters. We experimentally combine some criteria to cut the dendrogram: relative loss of inertia, cluster diameter and minimum number of clusters.

For next tests with SQLShare workload, dendrograms are not manipulable nor visually analysable. Then, based on the empirical results of [Moreau et al., 2021a], we considered the combination of UMAP [McInnes and Healy, 2018] and DBSCAN [Ester et al., 1996], that best performed on sequences of semantic elements. Preliminary tests with ground truth obtain comparable results for the artificial workload and improve results for the Ipums workloads [Moreau and Peralta, 2021]. Further experiments validate that DBSCAN is well suited to the topology resulting from CED and UMAP when applied on the SQLshare workload [Moreau and Peralta, 2021].

We test several strategies for selecting and normalizing query features, and setting CED, UMAP and clustering parameters; see [Moreau et al., 2020c, Moreau et al., 2022] for details.

Clustering quality. We first test our method with CED on the Artificial, Ipums, Open and Security workloads, obtaining respectively 4, 4, 6 and 5 clusters. In addition, on the Artificial and Ipums workloads (which have a ground truth), we compare with ED and AD measures.

Findings for CED. Clustering results using CED are reported in Table 2.14, and a visual comparison of clusters with ground truth (using dendrograms) is shown in Figure 2.3 (top part).

A first remark is that CED obtains good-quality results, and in particular, ARI and V-measure scores evidence a pure partition of the Artificial workload and reasonable partition of the Ipums workload. Indeed, dendrogram (a) exhibits a perfect match for the Artificial workload. We expected a good result with this workload, as CubeLoad templates are well differentiated. In addition, many explorations of the Slice All template (and some of the Slice and Drill template) are highly similar (distance near 0) as they contain sequences of the very same operations, even if exploration size is variable. This is one of the characteristics that makes CED a well-adapted distance for this problem.

CED also correctly classes most FOCUS and ATYPICAL explorations of the IPUMS workload (see dendrogram (d)). However, it fails to distinguish between OSCILLATE-FOCUS and OSCILLATE explorations, the frontier being quite fuzzy, and FIX explorations are not distinguished from FOCUS ones. We remind that these labels are not a real ground truth, but a preliminary classification for other purpose.

In addition to ARI and V-measure scores (calculated w.r.t. a ground truth), we compute cluster diameters and Silhouette scores to complete our quality analysis, for the 4 workloads (also reported in Table 2.14). Globally, we observe that most diameters are low, indicating that clusters are compact. Therefore, medoids are good representatives of each cluster. Most Silhouette scores are also positive, which is a good result given that our clusters are not hyper-spherical. In particular, we note that even if CED was able to generate a pure partition for the Artificial workload, we observe a low Silhouette score.

[17] `https://github.com/ClementMoreau-UnivTours/CED_Dolap`

[18] `https://colab.research.google.com/drive/1b9L-45zd9CEgFF4_ux7fwBLc-R9k_2i6?usp=sharing`

Workload	Nb clusters	Max diameter	Silhouette	ARI	V-measure
Artificial	4	2.22	0.49	1	1
Ipums	4	3.31	0.28	0.29	0.42
Open	6	4.53	0.37		
Security	5	3.81	0.16		

Table 2.14: Nb of clusters, diameter, Silhouette, ARI and V-measure scores using CED

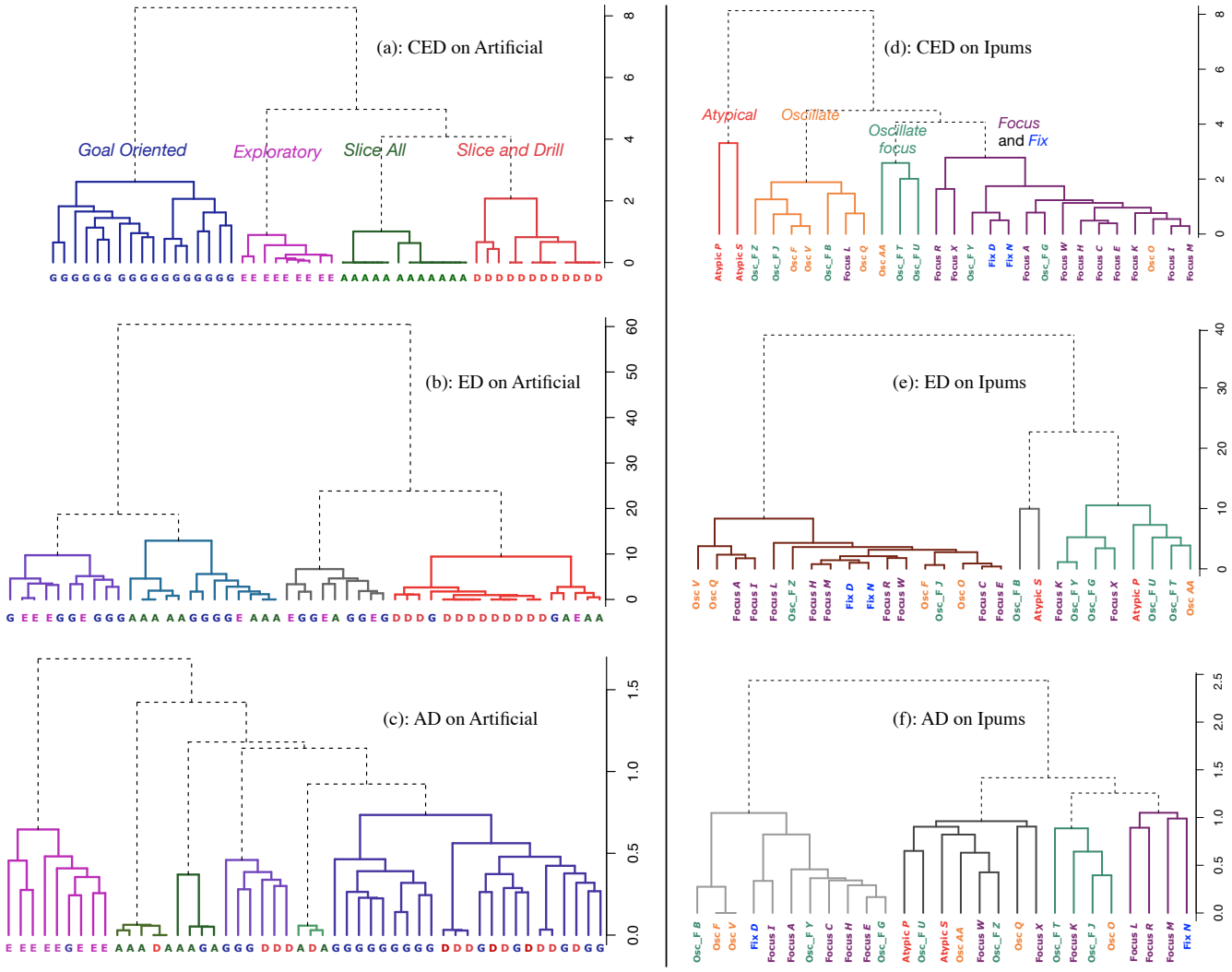


Figure 2.3: Dendrogram results on Artificial (on left) and Ipums (on right) workloads, with CED, ED and AD measures. Explorations are arranged in the horizontal axis, plotting similar explorations close. Links indicate which explorations are clustered together, shorter links meaning more similar explorations (vertical axis reports distances). Links of the same color represent a cluster, while dotted links just indicate inter-cluster distances. For easing the interpretation we also color explorations ids, according to ground truth labels. We deliberately chose the same set of colors as clusters to visually highlight the good matches.

Workload	Distance	Nb clusters	ARI	V-measure
Artificial	CED	4	1	1
	ED	4	0.26	0.36
	AD	6	0.76	0.88
Ipums	CED	4	0.29	0.42
	ED	3	0.09	0.23
	AD	4	0	0.19

Table 2.15: Comparison of clustering results for CED, ED and AD distances

Comparison to ED and AD. Table 2.15 reports quality measures for comparing the clusters obtained with CED, ED and AD measures. We remark that CED outperforms AD and ED in both Artificial and Ipums workloads, for both quality metrics.

With ED, clusters reflect exploration sizes instead of query operations. For example, in Figure 2.3(e) the first cluster includes short explorations, the second cluster contains the longest ones, and the last cluster contains medium ones. Conversely, AD relies more on the actual query parts to establish similarity, and tends to cluster together explorations navigating in the same portion of a data cube. Consequently, very different behaviors (e.g. those of Slice and Drill and Goal Oriented templates) are clustered together (see Figure 2.3(c)).

On the other hand CED is solely based on the structural properties of the explorations, which proves to be a good strategy.

Analysis patterns in OLAP explorations. An in-depth analysis of the obtained clusters, revels 11 analysis patterns, some of them being observed in several workloads. Table 2.16 summarizes the learned patterns, a detailed description, including observations for each workload, can be found in [Moreau et al., 2020c]. We highlight here our main findings.

We first remark that from the four CubeLoad templates, only the *Slice and Drill* pattern was found in real explorations.

As expected, analysts' behavior (in the Security workload) is different from students' (in the Ipums and Open workloads). Globally, analysts' explorations exhibit less operations, with more emphasis in the grouping of data, probably also in their arrangement and visualization (which is not captured in our method) of the data; while students are more click-oriented and produce longer explorations with much more operations.

In addition, our study shows that 40% of the students' explorations follow a focused pattern (*Slice and Drill*) translating that those students have developed a particular type of analysis skills, while 27% of students are more exploratory (*Oscilating*), perhaps translating a lack of maturity in their analysis skills, perhaps just showing their style.

From a more general perspective, the ability of distinguishing among skilled and novice users opens the door for more personalized EDA support. The recognition of outlier behavior is another strong point of our method.

Analysis patterns in SQL explorations. The third clustering variant (the one mixing operations and complexity features) does not succeed. We test several configurations, but in all tests, we obtaine well defined clusters for one-shot explorations and a small number (1 or 2) of huge mixed clusters for longer explorations.

Consequently, we only report here our findings for the second clustering variant. We obtain 12 clusters (6 clusters for one-shot explorations and 6 for longer ones).

Pattern	Description	Workloads
Slice and Drill	Focused explorations, continuously increasing the level of detail and filtering. Some of them mix other operations, mostly at the beginning, but drill-downs and filters are the predominant operations.	A,I,O
Slice All	Explorations with sequences of unfilter/filter operations.	A
Exploratory	Explorations with varied random operations.	A
Goal-oriented	Explorations with varied operations but converging to some specific point.	A
Oscilating	Explorations alternate drill-downs and roll-ups, oscillating the level of detail. Most of them do not filter any data.	I,O,S
Oscilate +Focus	Explorations start alternating drills down and rolls up, then alternate filters/unfilters, some of them focusing at the end.	I
Constant Agg. level	Many short explorations, with constant or lightly increasing level of detail. Most of them have little filters, but exhibit some changes in the measure set.	O,S
Add-Delete Fragment	Long explorations, with most queries alternating the addition and deletion of one fragment (a level in the group by set, a filter or a measure).	S
Few operations	Many short explorations, with few operations, mainly drill-downs.	S
Repeted queries	Explorations very few operations, globally exhibiting long subsequences of identical queries.	S
Outliers	Explorations clustered alone, with erratic behavior.	I,O,S

Table 2.16: Summary of discovered patterns and the workloads where they were found (A: Artificial, I: Ipums, O: Open, S: Security)

Clusters are analyzed with a variety of indicators (described in Subsection 2.5.2), evidencing 12 patterns of common or less-frequent behavior. The most prominent aspects of each pattern are summarized in Table 2.17, a detailed description can be found in [Moreau et al., 2022]. We highlight here our main findings.

Firstly, 49% of explorations are one-shot. They differentiate in the predominant operations in their single query. We discover 6 patterns. The most common one (C_1) consist in evaluating a simple query (with few complex clauses), projecting many attributes, possibly to verify it the dataset is correctly uploaded or just looking at the data.

Less frequent patterns, also concerning the evaluation of a simple query, differentiate in the used SQL operations, namely, many selections (C_2), many projections with some selections (C_3), aggregation and grouping (C_4), join of multiple tables and many set operations (C_5), and ordering (C_6). The latter is an outlier behavior, only concerning 19 explorations and concentrating most expert clauses. The usage of complex clauses is higher in clusters C_2 and C_6, while marginal in cluster C_4. These 5 patterns suggest a more specific analysis of data (w.r.t. the common behavior in C_1), taking advantage of more SQL operations and clauses. This may reflect users' preferences on some SQL clauses, but may also reflect users' expertise.

The remaining 51% of explorations contain between 2 and 98 queries, median being 4 queries. We discover 6 patterns, all concerning more complex clauses than those of one-shot explorations, in particular, advanced joins and advanced clauses.

A common pattern (D_1) reveals long explorations, with few operations per query, sometimes repeating queries, which translate a focused data analysis. Many types of operations are used, but mostly once per query, suggesting a conscious use of SQL.

	Cluster	Med. |exp|	Med. # op.	Common op.	Freq. agg. op. vect.	Complex clauses	Freq. agg. clause vect.	Workload coverage	Pattern nickname
One-Shot	C_1	1	6	+P +T	PT, PST	18%	V	35%	Full PROJECTIONS
	C_2	1	5	+P +S +T	PST	39%	J	3%	FILTER enthusiast
	C_3	1	5	+P +S +T	PST	19%	J	5%	Only PST
	C_4	1	7	+P +A +G	PAT, PATG	9%	V	3%	AGGREGATOR
	C_5	1	2	+P +T	PT	16%	U	2%	Table JOINER
	C_6	1	7	+P +T +O	PTO	37%	V, J, E	$< 1\%$	Ordered
Longer exp.	D_1	8	2	+P -P +S	$\emptyset$, S, P, PST	51%	J, V	16%	Long & Focused
	D_2	3	6	+P -P +T	P, PST, PS, PT	47%	J, V	19%	PROJECTION chains
	D_3	4	3	+P +S -P -S	S, PST, P	30%	J	7%	FILTER chains
	D_4	2	3	+P +T -T	T, PST, PT	32%	J, V, U	2%	Dataset reloader
	D_5	2	0	+P +T +S	$\emptyset$, PT, PST	34%	J, V	5%	Repeater
	D_6	3	3	+P +O +T	O	28%	J, V	2%	ORDER maniac

Table 2.17: Summary of learned patterns: median exploration length, median number of operations per query, common operations, frequent aggregated operation vectors, percentage of queries with complex clauses, frequent aggregated clause vectors and workload coverage

Another common pattern (D_2) reveals short explorations, with more operations per query. Projections are omnipresent, but frequently combined with other operations. What is interesting here, is the chaining of the same types of operations along the exploration. It can be exploited for providing personalized suggestions to users.

Two interesting but less frequent patterns (D_3 and D_5) concern a simple first query (with some projections, selections and joins), followed by chains of selections (D_3) or repeated queries (D_5). In both cases, explorations are shorter than in D_1 but reveal some kind of analysis. While D_3 suggest a meticulous study of the dataset, D_5 includes many novice users trying to understand how SQL works.

A similar but more complex pattern (D_6) involves richer first queries, followed by changes in the ordering of projected expressions. In addition to a good use of SQL, this behavior may correspond to users looking for the best way of reporting data.

The last pattern (D_4), also less frequent, exhibits a particular behavior. It concerns many changes in the datasets (frequently, the unique operation in the query is a change in the FROM clause). This corresponds to the upload of a new dataset and the execution of the same query on the new dataset, and suggests data analysts dealing with quality issues in their datasets.

Comments on users' skills. In order to study how particular users analyse data, we inspect how the explorations of each user are distributed across the clusters. To avoid noise, we discard 61 users having no cluster with at least 3 explorations (this includes 33 one-shot users). We observe three types of users: (i) 14 users whose explorations are concentrated in at most 3 clusters, (ii) 17 users whose explorations goes to more than 3 clusters but who made many explorations ($\geq$ 30) allowing the discovery of predominant clusters, and (iii) 5 users that made less explorations, distributed along many clusters. For 35 out of 36 such users, there is a cluster containing more than a quarter of user's explorations, and for 11 of them, there is a cluster containing more than a half of their explorations.

We complement these results by studying query complexity per user. In Table 2.18 we count the number of users that use (in some of their queries) explicit attributes (B), tables (T), sub-queries (Q), functions (F), advanced joins (J), set operators (U), advanced clauses (V), expert clauses (E), or any complex clauses CC; lines allow to distinguish users according to the total number of queries they made. We remark that 16 users made a single query, with basic clauses, 27 users made among 2 and 9 queries, with few complex clauses, 36 users made among 10 and 99 queries, using functions and some complex clauses, and 18 users made more than 100 queries, with greater level of complexity.

More generally, users explore data in several ways, with different operations and devising queries of diverse complexity. Interestingly, many users alternate among one-shot and longer explorations. Nevertheless, dominant patterns are visible for most users and this knowledge can be exploited by EDA tools.

		Users making use of								
Total # queries	**# Users**	**B**	**T**	**Q**	**F**	**J**	**U**	**V**	**E**	**CC**
$>$ 100	18	18	18	16	18	15	8	17	10	18
10-99	36	36	36	7	30	14	5	10	3	23
2-9	27	25	27	1	12	8	1	1	0	10
1	16	12	14	0	3	5	0	0	0	5

Table 2.18: Number of users using attributes (B), tables (T), sub-queries (Q), functions (F), advanced joins (J), set operators (U), advanced clauses (V), expert clauses (E), or any complex clauses (CC = J $\cup$ U $\cup$ V $\cup$ E), distinguishing according to the total number of queries they made.

2.5.4 Discussion

This section presented an original solution to learn analysis patterns in query workloads, which understanding has great implications for query recommendation, monitoring, optimization and, more generally, providing better EDA support.

The proposal includes an abstraction of queries and explorations in the space of query operations and complexity indicators, a set of similarity functions tailored for queries and explorations, and an innovative clustering process taking advantage of UMAP reduction for analysing a complex space.

We used a large palette of indicators for profiling the workload and analyzing the obtained clusters under several angles. These sets of statistical and visual indicators allowed to report how SQL operations are frequently combined and chained, and how complex are queries, both in terms of expressiveness and usage of complex clauses and functions. In addition, we analysed users according to their analysis behaviors.

The approach was tested on real workloads, allowing the extraction of 11 analysis patterns of OLAP explorations and 12 of SQL explorations. The former set includes 2 frequent patterns, corresponding to focused (Slice and Drill) and exploratory (Oscilating) behavior, but also evidenced other patterns used by experts and novice users. The latter set includes 3 typical behaviors: one-shot simple explorations, short exploratory explorations, and longer more focused ones, but also less-frequent behavior evidencing the punctual use or the chaining of specific SQL operations and clauses.

We believe that the identification of such behavior should be at the kernel of more intelligent EDA tools. In particular, such knowledge can be used for dynamically (at query time) classifying new users according to their behavior and analysis skills, and suggesting appropriate operations and clauses to complete their queries and continue their explorations. We believe that learning specific behavior and analysis needs of users is capital for developing EDA tools that go further than collaborative filtering. Indeed, a better understanding of users' navigation through the data can help to better understand their intentions but also their limitations. As a consequence, we hope that our work can help in a recommendation purpose of surprising and relevant queries for the user.

Finally, our proposal could be used on other workloads, specially those including queries generated by bots, as SDSS. Authors of [Singh et al., 2007] acknowledge the difficulty of extracting human sessions from all those collected: *"We failed to find clear ways to segment user populations. [...] Interactive human users were 51% of the sessions, 41% of the Web traffic and 10% of the SQL traffic. We cannot be sure of those numbers because we did not find a very reliable way of classifying bots vs mortals."* Developping tools helping in the recognition and analysis of hand-written queries is a nice challenge.

2.6 Conclusion

This chapter presented our contributions for understanding, modeling and learning the way users analyse data.

We first proposed a model for queries and explorations from the prism of users' skills. The model is based on a large set of features describing varied aspects of a query, including query text and result, but also its context within the exploration, including common query fragments, operations and timing. With the challenge of capturing incremental knowledge and regular behavior from a sequence of queries, we paid special attention to query operations and complexity.

We then proposed an approach for qualifying OLAP queries and explorations, modeled respectively as classification and skill acquisition problems, which succeeded to capture the characteristics of the workload, expert's advice and users' skills.

The extension from OLAP to a more complex SQL environment, and avoiding the need of experts for labeling, introduced new challenges. Consequently, we proposed an approach for workload segmentation and we studied query complexity issues. Finally, we proposed a similarity measure tailored for explorations, which paired with a clustering algorithm, allowed the identification of several types of analysis patterns.

Table 2.19 summarizes our contributions in terms of query model, query features, used workloads and main lessons learned.

Contribution	**Proposed models**	**Query language**	**Query features**	**Used workloads**	**Main lessons learned**
Exploration quality	Focus	OLAP-like	8 intrinsic, 3 relative, 8 contextual	Open	Good quality. Impact of all categories of features.
	Contribution		8 intrinsic, 9 relative, 8 contextual	Open, Enterprise	Good quality. Sensible to workload and labelling.
	Users' skills				Good quality. Good skill distinction.
Workload segmentation	Voting	OLAP-like, SQL	6 intrinsic, 6 relative	Open, Entrprise, SQLShare	Good quality. Outperforms state-of-the-art strategy.
	Transfer				
	Weak-labelling				
Analysis patterns	OLAP patterns	OLAP-like	2 intrinsic, 6 relative	Ipums, Open, Security, Artificial	Good quality. Good skill distinction. Outperforms state-of-the-art measures. Interesting patterns.
	SQL patterns	SQL	9 intrinsic, 12 relative	SQLShare	Interesting patterns. Relation with users' skills.

Table 2.19: Summary of contributions

In addition to the applications discussed in this chapter, our techniques have been used for analyzing other types of users' behavior. Indeed, our representation of explorations as sequences of queries, can be easily translated to other types of human activities, representable as sequences of complex elements. In particular, we studied:

– Human mobility, represented as sequences of semantic activities, such as walking, shopping or going to cinema (PhD thesis of Clement Moreau) [Moreau, 2021]. We studied intrinsic properties of sequences of complex elements and proposed several distances tailored to such sequences [Moreau et al., 2020b, Moreau et al., 2021a, Moreau et al., 2021b]. A general framework, and a web tool, SIMBA[19], were developed for managing, profiling, clustering and visualizing these sequences.

[19]SIMBA application: https://github.com/Clement-Moreau-Info/SIMBA

- Vehicle mobility, represented as sequences of moves and stops (PhD thesis of Frederick Bisone) [Bisone, 2021]. We studied raw trajectories of connected vehicles, concerning GPS and several sensors. We proposed a trajectory mining process for strong semantic enhancement of raw trajectories, allowing the identification of high-level semantic activities of vehicles during travel and stops [Bisone et al., 2019]. This process was instantiated to analyze the trajectories of connected ambulances of the Fire Department of Tours city in France.
- Writing skills, represented as sequences of books reviews (master project). We modeled the procedural knowledge needed to write helpful reviews, based on various metrics stemming from text analysis (like readability, polarity, spelling errors or length). We used Knowledge Tracing to measure the evolution of the ability to write reviews of good quality over a period of time [Megasari et al., 2018].
- Patient records, represented as sequences of medical appointments, with clinical and biological results (PhD thesis of Guillaume Tejedor, started in November 2023). We studied medical records of patients suffering from Amyotrophic Lateral Sclerosis (ALS) disease and identified potential similarity measures for comparing such sequences. We aim to cluster patients with similar disease evolution and predict survival time. Our first results are promising [Tejedor et al., 2024].

Our techniques have also been considered for other usages. In particular, CED distance has been studied in several contexts, for example, for pattern-driven analysis of pedestrian movement [Ali, 2022], semantic analysis of collections at the National Library of France [Zreik, 2023], clustering of sequence-based time use data [Becker, 2022], personal lifelong pathway co-construction [Ringuet et al., 2022], and analysis of sentence similarity [Wang and Ma, 2022].

Our benchmark has also been studied in the context of data quality alerts in Big Data analytics [Gyulgyulyan et al., 2019], and interactive data exploration of distributed raw files [Álvarez-Ayllón et al., 2019]

Next chapter investigates another important aspect for improving EDA tools, that is the consideration of users' interests.

Chapter 3

Understanding users' interests

This chapter describes our contributions for understanding, modeling and enhancing users' interests.

It relies on materials published in several conferences and journals, the main ones being [Drushku et al., 2019, Gkitsakis et al., 2024]. The overall contributions were developed in collaboration with several PhD and master students and colleagues, as summarized below.

Advising, projects and collaborations

PhD thesis:
Krista Drushku (2016-2019), *User Intent based Recommendation for Modern BI Systems*, co-supervised with Nicolas Labroche and Patrick Marcel.

Master theses and projects: Alexandre Chanson (2019), Antoine Chedin (2019), Ben Crulis (2019).

Other collaborations:
Panos Vassiliadis and Dimos Gkitsakis (University of Ioannina, Greece),
Stefano Rizzi and Matteo Francia (University of Bologna, Italy),
Alexis Naibo and Bruno Dumant (SAP Labs France, France).

Contents

3.1 Problems and positioning

BI users range from executives to data enthusiasts who share a common way of interaction, i.e., they navigate large datasets by means of sequences of analytical queries elaborated through user-friendly interfaces. For example, users may express their information needs via keywords, and let the system infer from them the most likely formal queries (generally MDX or SQL) to be sent to the underlying data sources (generally data warehouses or databases).

It usually takes many interactions with the system to satisfy an information need, and the overall session is often a tedious process, especially in the case when the information need is not even clear for the user. This bears resemblance with Web Search, where users typically need to repeatedly query the search engine to determine whether there is interesting content.

Being able to automatically identify users' interests from BI interactions is a challenging problem that has many potential applications, such as highlighting of the most interesting data, suggestion of interesting data found by other users, repetitive task prediction, alert raising, etc. that would help reduce the tediousness of the analysis. We are particularly interested in leveraging users' interests to devise recommendations of interesting queries, helping the users to pursue their interaction with the BI system.

Our research challenge is to **model query interestingness, learn users' interests** and take advantage for improving users' analysis experience.

Several research needs stem from this general challenge. They are described in the following subsections.

3.1.1 Need for modeling query interestingness

While many recent works propose techniques to assist query formulation for EDA [Idreos et al., 2015], as discussed in previous chapter, few works address the issue of leveraging users' interests without asking for explicit feedback.

A first problem to investigate is **what are the fundamental characteristics that make a query interesting for a user**, and therefore, **how to model and assess the interestingness of a query?**

Various interestingness measures were proposed in several areas of data exploration. For instance, there is a long discussion about interestingness in the area of evaluating recommender systems [Herlocker et al., 2004, Gunawardana and Shani, 2009, Kaminskas and Bridge, 2017]. We mention [Kaminskas and Bridge, 2017] as an excellent survey on the topic, discussing 4 criteria (diversity, serendipity, novelty, and coverage), in addition to the traditional accuracy, for evaluating the quality of a recommendation.

Many measures were proposed for pattern mining [Geng and Hamilton, 2006, Yao et al., 2006, Bie, 2013, Crémilleux et al., 2018]. Authors of [Geng and Hamilton, 2006], point out that interestingness is a broad concept and review 9 criteria to determine whether or not a pattern is interesting: conciseness, generality/coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability/applicability. Interestingness measures are also studied for the selection of formal concepts [Kuznetsov and Makhalova, 2018, Ibrahim et al., 2021].

Started with the seminal proposal in [Sarawagi et al., 1998], various interestingness criteria have been proposed to qualify an interesting property or pattern for a subset of the data in a dataset, often called insight, highlight, finding, discovery, etc., typically characterized by an interestingness score [Sarawagi, 2000, Gkesoulis et al., 2015, Wang et al., 2020, Bar El et al., 2020, Milo and Somech, 2020].

Despite the fact that many interestingness measures have been proposed in the literature, proposals do not follow a basic-principles approach, starting from the fundamentals of interest, to establish the ground for modeling measures.

Our goal is to **characterize what makes a query interesting** and take advantage for **facilitating the development of appropriate measures**.

Our research track is a two-level modeling. At the first level, we discuss *high-level aspects* of interestingness, deduced from the study of human behavior in modern philosophy. Moreover, we provide a structured taxonomy of how the analysts' goals, beliefs and intents as well as the computational environment, relate to the evaluation of the different aspects of interestingness. At the second level, we provide *data-oriented measures* of interestingness, substantiating the aforementioned high-level aspects, on the grounds of the available information.

Section 3.2 presents our contributions for modeling interestingness.

3.1.2 Need for discovery of users' interests

Having a characterisation of interestingness, and mechanisms for evaluating it, the natural next challenge is to discover users' interests, and then aid the user to devise interesting queries.

A first problem to investigate is **whether users' interests can be detected in a query workload**, and then, **whether such interests may help in the recommendation of queries.**

Many approaches analysing query workloads are more focused on extracting query patterns than users' interests, with the purpose of auto-completing SQL queries on the fly (e.g. [Khoussainova et al., 2010]) or pre-fetching queries (e.g. [Jayachandran et al., 2014, Sapia, 2000]). These works influenced later work on query recommendation (e.g. [Aufaure et al., 2013b, Eirinaki et al., 2014, Aligon et al., 2015]), aiming to find the most likely query to follow a given current query. A survey of query recommendation approaches is proposed in [Aligon et al., 2015], showing that users' interests are poorly considered in state-of-the-art.

The closest proposal to interest detection is an approach for discovering the most accessed areas of a relational database [Nguyen et al., 2015]. Their notion of user interest relies on the set of tuples that are more frequently accessed, and is expressed as selection queries (mostly range queries). They use DBSCAN to cluster user interests and use a similarity measure based on the overlapping of selection predicates.

Alternative ideas come from the Information Retrieval community, where the analysis of web search sessions for personalizing users' experience has attracted much attention [Mobasher, 2007]. Various forms of user interests have been defined, such as contextual intent [Sun et al., 2016], task repetition [Song and Guo, 2016] or long term interests [Guha et al., 2015], and methods have been proposed to identify them.

We highlight a proposal for discovering new intents and obtain content relevant to users' long-term interests [Guha et al., 2015]. Authors develop a classifier to determine whether two search queries address the same information need. This is formalized as an agglomerative clustering problem for which a similarity measure is learned over a set of descriptive features (the stemmed query words, top 10 web results for the queries, the stemmed words in the titles of clicked URL, etc.). One advantage of this approach is that it allows for the building of contexts that span over several users' sessions or only a portion of one session. Thus, contexts provide insights on short and long term information needs and users' habits, to build accurate user profiles.

Inspired by the work Guha et al. did in the context of Web Search [Guha et al., 2015], our goal is to **detect users' interests** based on users' past queries and to **leverage such interests for recommending queries.** The challenge of user interest detection lies in the fact that

interests are hidden in the interactions, and two users with the same interest would probably interact with the system differently. As in Web Search where users may have no idea of the retrieval algorithm, BI users are generally ignorant of the data sources and the formal queries they trigger. However, once logged, all this information (keywords, sources, formal queries, etc.) provides a rich basis for discovering user interests.

Our research track is to characterize user interests by means of features extracted from users' traces and group queries related to the same interests. We first use classification to learn a similarity measure that basically assigns a weight to each of the features. Then, we use such measure with an off-the-shelf clustering algorithm to group the queries. To leverage the discovered user interests for the purpose of query recommendation, we propose an original interest-based recommender.

Section 3.3 presents our contributions for interest discovery and interest-driven query recommendation.

3.1.3 Scope

We place on a particular case of EDA that is OLAP analysis of multidimensional data. As motivated in previous chapter, the analysis of OLAP workloads allows to qualify queries and learn users' interests in a relevant, simple, focused and rich environment.

Road map. Sections 3.2 and 3.3 present our contributions to the previously described research needs and Section 3.4 draws our conclusions.

3.2 Modeling interestingness

Foreword

This section summarizes some works in collaboration with Patrick Marcel and Panos Vassiliadis. In particular, recent work is related to the (ongoing) PhD work of Dimos Gkitsakis, and the master projects of Spyridon Kaloudis and Eirini Mouselli, all of them supervised by Panos Vassiliadis.
The proposal was published at ADBIS [Marcel et al., 2019] and DOLAP [Gkitsakis et al., 2023]. An extended version is in print in Information Systems [Gkitsakis et al., 2024].

How interesting is a piece of data?

In this section, we frame an answer to this question from the viewpoint of the study of human behavior. To the best of our knowledge, there is no formal definition of interestingness. Online resources[1] propose "Interest is a feeling or emotion that causes attention to focus on an object, event, or process". In contemporary psychology of interest, the term is used as a general concept that may encompass other more specific psychological terms, such as *curiosity* [Litman, 2005] and to a much lesser degree *surprise* [Reisenzein et al., 2012] and *novelty* [Förster et al., 2010]. From our study of the literature, we can conclude that interestingness is a degree attributed to a piece of information, regarding the curiosity and surprise it generates. This piece of information under consideration may spark the will to continue exploring the source of information to close some knowledge gap and explain peculiarities, or get novel information.

In order to pass from such a high level description of interestingness, to a more concrete one, our approach is a two level modeling. At the first level, we discuss *high-level aspects* of interestingness, deduced from the study of human behavior. Second, we provide *data-oriented measures* of interestingness, substantiating the aforementioned high-level aspects, on the grounds of the available information.

Considered workloads. We experiment on the Open, Adult and Loan workloads, described in Appendix A.

3.2.1 Interestingness aspects

In this subsection, we derive from our study of the literature the criteria of the interestingness of a piece of data by listing what influences them. To keep definitions simple, we focus in the interestingness of an individual piece of data (i.e. a cell in a data cube), as originally introduced in [Marcel et al., 2019]. But we remark that all definitions and measures are easily extendable to assess interestingness of a set of cells. In particular, in [Gkitsakis et al., 2024], we consider larger pieces of data, in particular, query results (i.e. query answers).

We present 4 fundamental, high-level interestingness aspects: relevance, novelty, surprise, and peculiarity.

Relevance as a measure for the user's curiosity. Curiosity is the main driver of knowledge acquisition. Data exploration, especially in an environment of Business Intelligence, is primarily related to the answering of an open question. So, it is realistic to assume that the user comes with a question for a particular subset of the multidimensional space, and the user's exploration

[1] `https://en.wikipedia.org/wiki/Interest_(emotion)`

has to do with "a walk" within this sub-space in order to answer the question. We call the aspect of interestingness that pertains to curiosity as the *relevance* of the cell with respect to the exploration and its underlying user's intention.

The main force, thus, of the assessment of relevance is the modeling of the users' intentions. Basically, we can discriminate between (a) the case where a description of the user's intention is given vs. (b) the case where no such knowledge is available. In the former, we deal with an expression of the user's interest as the space of a user's intention. In the latter, we need to learn the user's intention from the history of past activity, which, in turn, relies on the availability of the coordinates of the cells of the queries in the exploration and the schema of the cube.

Novelty. Novelty is also an aspect of interestingness that mainly pertains to the need of users to learn information previously unknown. The simple reporting of data that have not been previously reported might increase their interestingness.

The main force behind novelty is the existence of a history. A lesser influence is the availability of query answers (cell coordinates are sufficient to understand if the cell has never been seen). Without the knowledge of the history of the user's queries, novelty is practically a wild guess. When dealing with novelty, we are not primarily interested in the intention of the user, although it can affect the attention that a user pays to a particular cell (in other words, we assume all cells being equally probable to have been observed by the user).

Surprise. Not surprisingly, surprise is a major aspect of interestingness. Surprise occurs when our previous beliefs are disconfirmed or contradicted. This can happen either directly, when the expected value of an event proves to be significantly different than the actual value, or implicitly, when the disconfirmation of a certain fact deduces the disconfirmation of a dependent fact.

Clearly, the main prerequisite for evaluating surprise is the existence of a previous belief of the user. Without the existence of a structured model for the estimation of the previous beliefs, the assessment of surprise is impossible; for this case, it is only possible to measure some objective peculiarity intrinsic to the data (see below). Surprise can be measured using models leveraging the history of the user with the datacube, for instance to estimate beliefs.

Peculiarity. Peculiar information (i.e. differing in some way from other information) awaken curiosity and typically ask for further knowledge acquisition and explanations. Indeed, peculiarity is an important aspect of interestingness, being the corner stone for discovery-driven exploration [Sarawagi et al., 1998]. Consistently with the literature on datacubes, we use peculiarity to denote an intrinsic property of the data, i.e., the cell's value, when considered together with other related cells.

Peculiarity of a cell cannot be assessed in vacuum. Most typically, it can be assessed against the cells of the same query. Taken to extremes, it can also be evaluated by comparing the cell to all the previous cells of the history of the exploration – or even, to all the cells of the full history of the user with the datacube, i.e., including past explorations. Finally, peculiarity may also be calculated with respect to the unseen cells of the cube. The full instance, i.e., with measure values, of cells considered are prerequisites for this criterion.

Based on these 4 fundamental high-level interestingness aspects, we define interestingness of a cell as a vector of scores, defined over a set of interestingness measures.

Definition 3.1 (Cell interestingness) *Given a user's exploration over a datacube, the interestingness of a cell of this exploration is a tuple of scores for a list of interestingness measures, concerning the following aspects:*

- *Relevance: the extent to which a cell is related to the overall analysis intention of the user.*
- *Surprise: the extent to which a cell contradicts and revises the user's prior beliefs.*
- *Novelty: the extent to which a cell presented to the user is new, and previously unseen.*
- *Peculiarity: the extent to which a cell is different, and not in accordance with other cells presented to the user.* □

We intentionally do not differentiate between high-level and data-oriented measures. We support an extensible approach towards which measures would an interestingness assessment tool include. Next subsections develop this approach.

3.2.2 A taxonomy for the assessment of interestingness

In previous subsection, we identified three forces affecting the assessment of interestingness. The first one is the *informational context* around the data that we want to score. As a cell never comes alone, other cells provide context for assessing "how interesting is this cell, after all". The second force is how the very *nature of information* impacts interestingness. Indeed, interestingness defined using only the multidimensional schema can be quite different than defined when the full data come into play. The third force has to do with how interestingness is impacted by *what is known about the user*.

We now propose a taxonomy to categorize interestingness measures into 3 dimensions (using the 3 forces identified above), and for each dimension, we identify categories in an order of increasing available information to the interestingness assessment system.

Dimension 1: Data context. This dimension concerns the breadth of context over which we compare a cell, in order to compute its interestingness. Precisely, it is about what information to consider to define interestingness. We consider the following categories:

- *Cell*, i.e. the information under consideration. Only the cell is needed to compute its score, without any context.
- *Query result*, i.e., the information that can be examined simultaneously with the cell. The score is computed by comparing the cell to the other cells of the query result.
- *History*, i.e., the information already examined. The score is computed by also comparing the cell to cells previously seen by the user.
- *Cube*, i.e., the source of information. The score is computed by comparing the cell to other cells of the datacube (whole datacube or some subsets).

Dimension 2: Datacube structure. This dimension concerns the datacube structures that may be accessed to compute the interestingness of a cell. We consider 2 categories:

- *Schema.* Only schema information (e.g. dimensions, hierarchies, schema constraints) is available to compute the interestingness score. By slight abuse of notation, we will call schema of a cell its coordinates.
- *Schema+Instance.* The score is computed using information about both, schema and instance (i.e., measure values are considered in the instance of a cell).

Dimension 3: User model. This dimension concerns the knowledge about the user to be considered to define interestingness. We consider 3 categories:

- *Without model.* The score does not consider any knowledge about the user.
- *Factual models.* The score is computed by also considering knowledge about the user, typically represented by a user profile. Such knowledge may concern user's preferences and interests, position, typical tasks, and all static information describing the user.
- *Dynamic models.* The score is computed by comparing the cell to previous knowledge about the user and judged w.r.t. what was learned about they. Such knowledge may concern goals, topics of the task at hand, short-term interests, and in a more general perspective, all beliefs and lessons learned during data analysis. All this knowledge typically changes all along the analysis session. New queries serve to confirm (or contradict) previous hypothesis, making new ones, and consolidating conclusions.

Figure 3.1 summarizes the discussion by defining the space of available information and position interestingness aspects in this space.

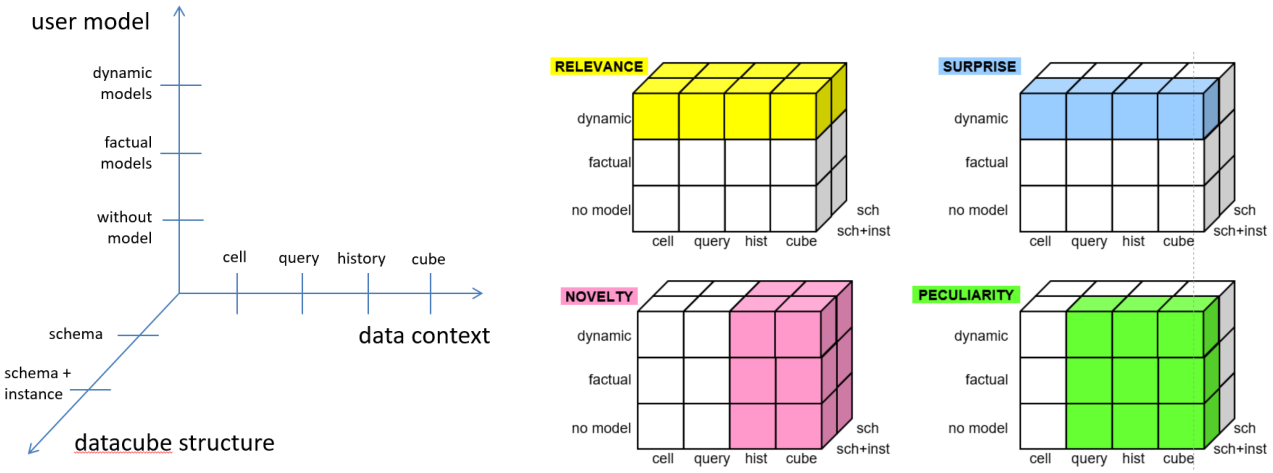


Figure 3.1: (i) Taxonomy of forces influencing the assessment of interestingness, and (ii) aspects of interestingness in the space of these forces.

3.2.3 Interestingness measures

In this subsection, armed with the tools of the previous subsections, we discuss how to compute specific measures for assessing the interestingness of a cell. As an example, and without trying to be exhaustive, we provide some alternatives for assessing *relevance*. Examples for the other interestingness aspects are discussed in [Marcel et al., 2019]. A more detailed discussion, including a large set of assessment algorithms, can be found in [Gkitsakis et al., 2024].

Assessing the relevance of a cell practically answers the question: *how close is this cell to the subset of the multidimensional space that the user intends to explore*? Two fundamental notions hide behind this formulation of the problem, the specification of an area of interest and the understanding of the user's intention.

We define the *area of interest* of a user's intention as the framing of a subspace of the multidimensional space (either intensionally via selection predicates, or extensionally, as a set of cells) for which the user wants to obtain information.

Then, given a specific exploration, with a user's intention as its underlying motive, *relevance* can be computed as the degree to which the cell overlaps with the area of interest of the user's intention. We remark that such comparison can be based on the cube *schema* only, i.e. only comparing coordinates, or based on *schema + instance* (both categories of Dimension 2 of the taxonomy). Regarding Dimension 3, the modeling of the user's intention clearly corresponds to the *dynamic model* category.

We discriminate between (a) the case we have an explicit expression of the user's intention, and, (b) the case we have not such information, and thus, we have to derive a model of user's intention. Let us proceed in exploring both cases.

Relevance in the presence of knowledge of the user's intention Here, we do not discriminate between an intention induced by a user profile, or a deliberate expression of the intention by the user. In most cases, *the intention is expressed as a Boolean predicate* ϕ (typically -but not obligatorily- expressed as the conjunction of simple atomic selection formulae).

There are several ways to compute the relevance of a cell c to intention ϕ. Note that ϕ may not be part of the query that retrieves c. The user may (a) compare cells within the area of interest of the intention with similar/peer cells, or, (b) put the observed values in context by rolling-up in a way that produces aggregate values broader than the original area of interest.

The simplest way is to see whether c satisfies the intention ϕ. To do that, both c and ϕ must be converted to the same level of detail, typically, to their highest common descendant in the lattice of group-by's [Harinarayan et al., 1996], or the lowest possible node of the group-by lattice, i.e., the level of the facts, that we call C^0. Then, *relevance* in its simplest form is Boolean and evaluates to true if all descendants of c satisfy ϕ, or numerical, if a percentage is computed.

In these variants, the history of queries is not taken into consideration –only the intensional area of interest of the user's intention.

Relevance without knowledge of the user's intention. In the case where no model for the user's intention is given a priori, we can emulate it as the portion of the cube that has been more visited during the exploration.

To assess the relevance of a cell, we need to quantify how "close" or "central" the cell is to the subspace induced by the exploration of the user. Practically speaking, we need an algorithm that enumerates the cells that have been visited by the user during the exploration. Due to the hierarchical nature of the space, the easiest way to compare cells is by referring all cells to a common level of granularity.

Then, we need an algorithm that computes the area of interest S^0 at the level of C^0. The input to this algorithm is the history of user's queries of an exploration. The output is the detailed area of interest. Basically, for every aggregate cell that is part of a query result, the algorithm detects its detailed cells, increases a score for each of the times this cell has contributed to the computation of a query result and adds it to the detailed area of interest, returned by the algorithm. An alternative concerns finding a most concise description of S^0 by rolling up regions of C^0 completely covered by cuboids at an ancestor level at the lattice of group-by's.

Now, we can compute the relevance of a cell c to the computed area of interest S^0, as a function f_R that calculates the percentage of descendants of c at C^0 that also lies within S^0.

Variants. A more liberal definition of relevance can compute a distance function of the two sets. A more stricter definition might take the frequency of the visits of the user to each member of S^0 during the exploration. Then, each cell is weighted by how many times it has been visited by the user during the exploration. In this case, *relevance* is computed as the fraction of the sum of the weights of the common cells of the two sets over the sum of weights of the cells of S^0.

The algorithms mentioned in this section are described in [Marcel et al., 2019]. A large set of alternative algorithms are discussed in [Gkitsakis et al., 2024]. We include result-based algorithms, executed after query evaluation, and syntax-based algorithms, executed before query evaluation. The former set is used for highlighting interesting cells in the query answer, the latter for recommending interesting queries.

3.2.4 Experiments and results

In this section we report the major findings of our experiments for measuring query interestingness. In what follows, we consider the Open, Adult and Loan workloads. The former is used for measuring the interestingness of real users' queries. The short explorations of the Adult workload were especially devised to finely observe users' preferences in a controlled environment. The Loan workload is used for scalability tests.

Protocol. Our first experiments analyse the interestingness of users' queries in the Open workload, which is labeled according to query focus and exploration quality. Our goal is to confront the interestingness measures with the labels assigned to the exploration and queries, looking for correlations between interestingness, query focus, and exploration quality.

To this end, we compute 4 basic interestingness measures, one per high level aspect:

- simple relevance, without knowledge of the user's intention, computing the area of interest (as described in Subsection 3.2.3),
- binary novelty, i.e., the cell has been previously seen or not,
- a limited form of surprise, called positional surprise, computed as minus log of the product of the member's probability of appearance in the user's history, and,
- simple peculiarity, called outlierness, calculated as z-score w.r.t. the rest of the cells in the query result to which it belongs.

We also conduct a user study (detailed in Appendix A), in order to evaluate how do the interestingness aspects relate to the behavior of people working with cubes and cube queries. In particular, we investigate whether there are significant influences of the interestingness aspects in the users' perception of interest. The 25 participants were asked to *assess how interesting a query result appeared to them* based on their personal criteria; they were agnostic of interestingness aspects. Participants received 3 sets of 4 queries to be ranked according to their interest (from 1=most interesting to 4=less interesting, without ties) and were asked to justify their ranks.

The trick, unknown to the participants is that each of the 4 queries maximizes the value of an interestingness aspect, i.e. there are a highly relevant, a highly novel, a highly peculiar and a highly surprising query at a random order. *Thus, by ranking queries, the participants also ranked interestingness aspects without knowing.* As in the first experiment, we select 4 interesting measures, one per aspect.

Another experiment tests the scalability of the algorithms along three tunable parameters: (a) cube size, reflecting the number of tuples in the fact table, specifically: 100,000, 1 million, and 10 million tuples, (b) result size, reflecting the number of tuples in the query result, specifically: 10, 84, and 792 tuples, and, (c) history size, i.e. the number of the user's previous queries, i.e., specifically, 1, 5 or 10 past queries. The goal of this experiment is to assess the efficiency of the algorithms, via their execution time, by tuning the scale of the parameters.

Implementation and setting. In order to relate interestingness and exploration quality, we developed a prototype that loads the explorations of each user, and for each of them evaluates the queries one by one, in order. Each time a query is evaluated, the user's history is updated, the

detailed area of interest (as explained in Subsection 3.2.3) is refreshed and the cell interestingness measures are computed. Our prototype is written in Java 8 and ran on a MacBook Pro Core I5 with 16GB RAM running MacOS Mojave 10.14.3.

Scalability experiments are performed in a server with an AMD Ryzen 9 5900HS 3.3GHz CPU processor, 16GB of RAM and a 1TB SSD NVMe M2 hard drive. 8GB of RAM is allocated to the MYSQL server, via Workbench 8.0 CE.

Interestingness relation to exploration quality. We investigate whether the explorations with more focused queries and higher quality obtain higher values for interestingness measures.

Relation with focus. The first results come from Table 3.1. We average interestingness values of all focused vs non-focused cells. The focused category consistently demonstrates higher values for all the measures, with novelty having a 15% difference in the values and relevance a 10%, even though this is nuanced by the standard deviation.

	Relevance	Novelty	Surprise	Peculiarity
Not focused	0.68 (0.43)	0.56 (0.50)	0.77 (0.25)	0.61 (0.90)
Focused	0.78 (0.31)	0.71 (0.46)	0.82 (0.26)	0.66 (0.78)

Table 3.1: Average and Standard deviation (in brackets) of measures per query labels

Then, we refine the above result by assessing whether there is any difference in the behavior of these measures during the progression of the explorations. As exploration lengths are different, for each query we compute the percentage of progress with respect to the exploration, as an indicator of how deep the analyst was during that exploration. To reduce the visual clutter, we organize the demonstration by ranges of 10 steps, where the average value is shown for each category. Figure 3.2 shows how the four measures evolve along the progression of the explorations, distinguishing by query labels.

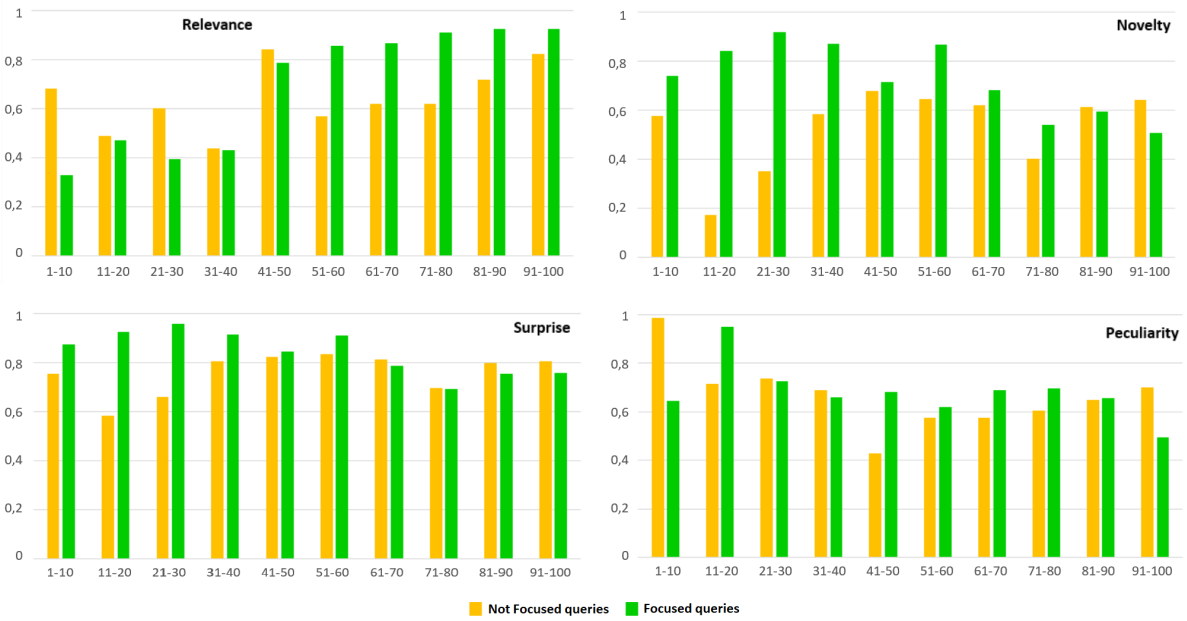


Figure 3.2: Evolution of the four interestingness measures (y-axis) with respect to the % progress in a exploration (x-axis) for focused vs non-focused queries

Concerning novelty, we see that focused queries soon demonstrate higher amounts of novelty compared to non-focused ones (which seem to revolve around the same cells). Only very later in the exploration is this difference equalized or surpassed (and indeed at low levels of novelty anyway). So overall, focused queries demonstrate more novelty than the non-focused ones. The same phenomenon is observed for surprise, but with less variations.

For relevance, as expected, we observe lower values at the beginning of the exploration, where users' intentions are less defined. Interestingly, non-focused queries, due to their repetition, obtain higher values than focused ones. Only later in the exploration, when the focused queries are returning to the well-established area of interest to finalize conclusions is the situation reversed.

For peculiarity, things are pretty much equal throughout the entire exploration, apart from a few cases where focused queries contain a little bit more outlier cells than non-focused ones. This justifies the small 5% advantage they have in the total scoring of Table 3.1.

Relation with exploration quality. Figure 3.3 shows how the four measures evolve along the progression of the exploration arranged by exploration label (ranging from A = high quality, to C = low quality, as explained in Section 2.3). The following general behaviors can be observed:

- C explorations are erratic, and novelty is low, one could say that users are not really analyzing, in that they are merely comparing with novel facts.
- In B explorations, all measures are high, there is too much movement, indicating that users are focused, but not enough. The fact that novelty and relevance are high at the same time is not contradictory: users stay in the same detailed area, but keep rolling-up, drilling-down. In other words, they keep investigating, but seem inconclusive, which is corroborated by the fact that those explorations are often longer than A explorations, that get straight to the point, and also by the fact that peculiarity tends to increase in the end.
- In A explorations, relevance keeps increasing, novelty is high then collapses, like surprise, and then start increasing again. This indicates that the explorations are more focused in the end. Peculiarity is very high in the beginning, which could have sparked the exploration.

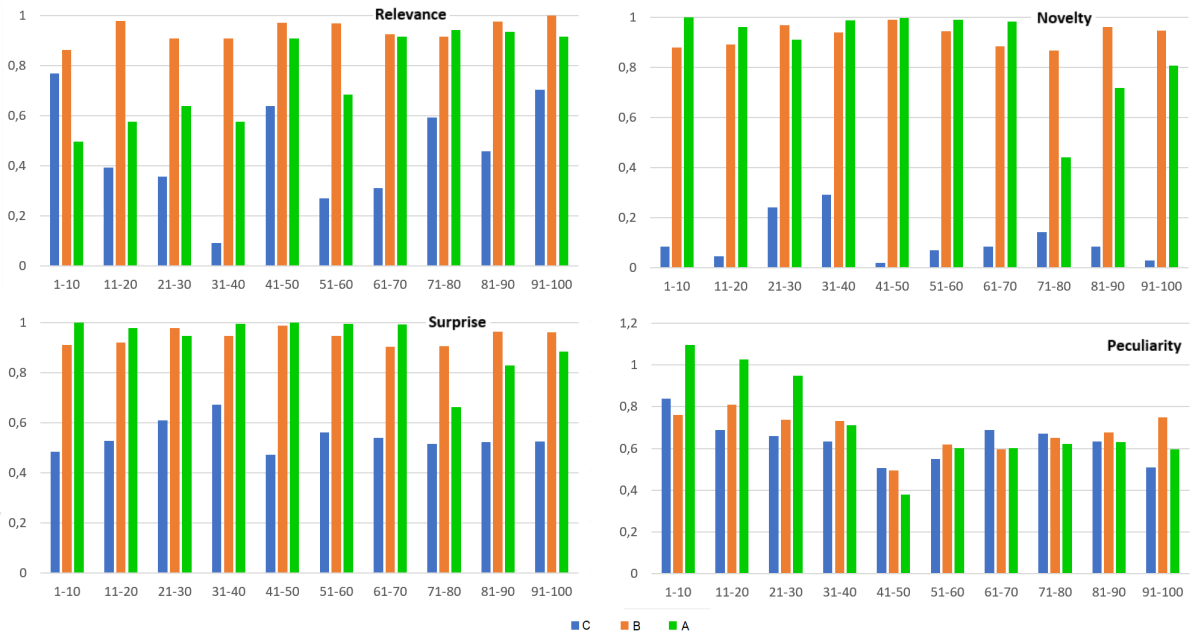


Figure 3.3: Evolution of the four interestingness measures (y-axis) with respect to the % progress in a exploration (x-axis) for exploration labels

Influence of interestingness aspects. As a result of the user study we collected the ranks (from 1 to 4) of each participant, for each query. Remember that ranking queries translates to ranking interestingness aspects. In order to investigate whether there is any dominant interestingness aspect, we compute Borda scores[2] for ranks, and total scores per interesting aspect. We highlight here our main findings.

Significance of individual interestingness aspects. Borda totals, shown in Table 3.2 suggest that no particular interestingness aspect drives the overall interest single-handedly. However, there are differences, with surprise and relevance being most significant, novelty coming third at a distance, and peculiarity being the least significant. We remark that surprise is also the aspect being most ranked 1st and less ranked 4th. Closely following surprise, relevance ranks typically 1st or 2nd, and less frequently 3rd or 4th. On the other hand, novelty is practically equally distributed in all ranks. Finally, peculiarity goes particularly low in terms of preferences. This practically instructs us that if recommending queries to users, surprise and relevance seem stable choices.

Interestingly, a statistical analysis of aspects correlation (with pairwise Pearson correlation) finds a couple of interesting anti-correlations: Surprise is anticorrelated with novelty, with a score of -0.62 and relevance is anticorrelated with peculiarity with a score of -0.50. The effect for the other pairs is weaker.

Aspect	**Borda score**	**Occurrences/rank**			
		1st	2nd	3rd	4th
Peculiarity	151	7	14	27	27
Novelty	183	20	16	16	23
Relevance	203	19	28	15	13
Surprise	213	29	17	17	12

Table 3.2: Total of Borda scores and occurrences per rank, for the interestingness aspects

Interest change over time To the extent that participants received 3 sets of queries to rank, ordered over time, we assess the effect of time via the position of the respective sets of queries (i.e. at the beginning, middle or end of the exploration). A caveat here is that the explorations are short, therefore, the results should not be arbitrarily generalized.

In Figure 3.4, we depict the average Borda score per position, for each of the interestingness aspects. Unsurprisingly, surprise and relevance seem rather unaffected from the position of the query in the exploration, although as time passes, surprise becomes slightly less of importance. Peculiarity also seems to lose interest as time passes, especially between beginning and middle queries. What is most revealing, though, is the sharp increase of novelty score over time. At the beginning, novelty is not that interesting, scoring lower among all interestingness aspects. Then, novelty starts being more appreciated by the participants. Novelty is probably considered out-of-scope at the beginning, but later, it picks up in stature.

Participants behavior We carry out other experiments in order to investigate whether participants demonstrated a consistent behavior in their rankings, and whether they could be clustered based on their preferences. To this end, we use several consistency measures (based on Boolean- and score-based differences) and test several clustering strategies (several algorithms, several features). Details can be found in [Gkitsakis et al., 2024]. We summarize here our results.

[2]Borda counting is a popular positional voting system, where scores (computed from votes) are totaled for designating the winner [El-Helaly, 2019]. Given N voting options, and a rank K, $1 \leq K \leq N$, a Borda count is computed as $score = N + 1 - K$.

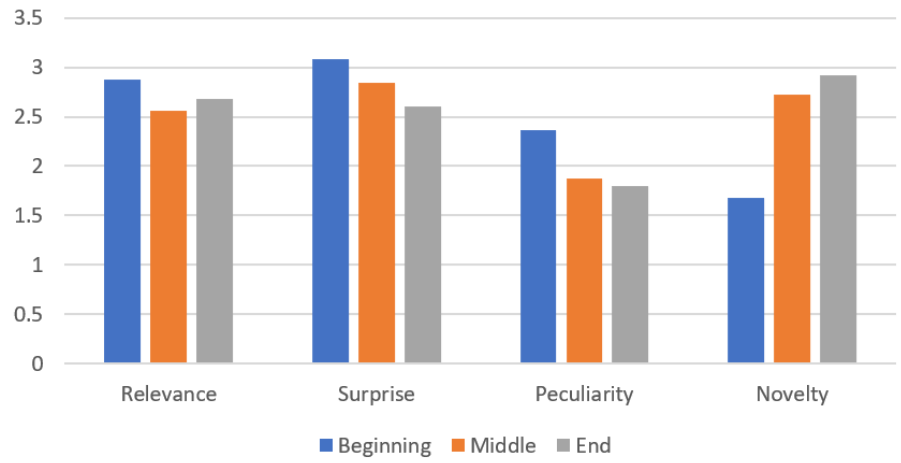


Figure 3.4: Average Borda scores of interestingness aspects per position

With the exception of a single participant (with a consistency of exactly 1), we find that more often than not, the rankings of the same interestingness aspect are different. But although ranks may not coincide exactly, they are not that far. In other words, the participants do not exhibit a strong bias towards a particular ranking of the interestingness aspects, although the rankings are not completely arbitrary. In addition, our results are quite indicative on the absence of clusters of participants. All clustering methods return low Silhouette coefficients, and their Silhouette plots indicate that clusters are not very cohesive. Interestingly enough, we do not observe any particular differentiation between the educational level of participants.

Scalability tests. In this experiment, we study the effect of the fact table size and the query history size to the execution time of 8 algorithms computing different interestingness measures. We test with 2 algorithms per interestingness aspect, with varied information in terms of the taxonomy dimensions.

Detailed results are reported in [Gkitsakis et al., 2024]. We simply highlight that most algorithms exhibit a linear increase of execution time w.r.t. the evaluated criteria, which agree with the complexity analysis of the algorithms. Nevertheless, there is one exception: An algorithm computing value-based surprise (comparing actual values in the query results with expected values in a user model), even being quite fast, do not achieve the theoretical lineal increase with respect to the result size. We attribute the variation to the probability of hitting an expected value when the result size is larger, which results in extra CPU time for computing surprise.

In absolute terms, of course, algorithms only basing on syntactic aspects (cube schema) run much faster than those taking advantage of instances, history or user models. In particular, the computation of detailed area is highly time consuming (almost 100 seconds for 10 million facts).

3.2.5 Discussion

This section addressed the problem of assessing the interestingness of the data analyzed by a user during data exploration.

We proposed criteria of interestingness at both, high level and data-oriented level. Indeed, we discussed 4 fundamental interestingness aspects: relevance, surprise, novelty, and peculiarity, and for each one, we proposed a large set of measures and algorithms for assessing them in a quantitative fashion.

We experimented with a workload of hand-written OLAP queries and conducted a user study, finding out that high-quality queries and explorations concern more interesting data and that there is no dominant interestingness aspect. Noticeably, users do not exhibit a strong preference on particular aspects and interesting perceptions change as time passes. Practically, relevance and surprise are generally more important for users, being stable along the exploration, while peculiarity has more impact at the beginning, and novelty later.

We believe that the assessment of interestingness aspects should be a first class citizen in EDA tools. On the one side, including assessment algorithms within query evaluation allows the highlighting of interesting values among the query results presented to the user. Many visualization and storytelling techniques may take advantage of interestingness measures and propose adapted visualizations. In [Chédin et al., 2020], we presented a proof of concept in this way. On the other side, assessment algorithms can be used offline, for assessing interestingness of past queries (as done in our first experiment described in Subsection 3.2.4) and enhancing query recommendation.

Among the large palette of algorithms proposed in [Gkitsakis et al., 2024], we took care to include both, result-based algorithms, to be executed after query evaluation and syntax-based algorithms, to be executed before query evaluation. However, it is clear that the proposed algorithms are only a first attack to the problem. More algorithms and metrics are possible for the aforementioned aspects, and more experiments are necessary before generalizing our findings. In particular, personal profiles, crowd-wisdom and log mining can be employed to best model users' beliefs. We refer the interested reader to [Bie, 2011, Bie, 2013] for a starting point, but of course, the problem of belief estimation is a large research territory that can fit gracefully with our taxonomical framework. In addition, the scope of our user study has not studied highly interactive user sessions. The extent that interactivity affects the assessment of interestingness is yet another unexplored territory for future research. The role of time (but also space) is also worth pursuing: what is interesting now for an analyst, might be indifferent some time later. Aging, decay factors can be introduced in the assessment of interestingness when queries are compared to the history of the user, or other users.

We have assumed a hierarchically-structured multidimensional space. Nevertheless, we have kept our discussion independent from the particular model of OLAP operations that can be applied to the data, or from technological aspects influencing it. We believe that our work opens the road for a more directed research of interestingness assessment and recommendation algorithms with specific targets among the high-level aspects discussed here.

Generalizing to a more complex SQL environment introduces new challenges, as was discussed in Chapter 2. This OLAP configuration also means that the data are clean and conforming to the designed hierarchies. We can imagine that the presence of arbitrary schema structures and arbitrary values in the data should highly impact interest perceptions, specially surprise, and may trigger the discovery of peculiar (wrong) data. Our experiments provide a proof of concept in this direction, showing how even simple measures can help the analysis of users' behavior. Extending the framework beyond the realm of clean, simply structured multidimensional spaces, in the realm of an arbitrarily structured and populated database schema, is a clear path for future work.

Next section illustrates the usage of interestingness for enhancing EDA.

3.3 Learning users' interests

Foreword

This section summarizes part of the PhD thesis of Krista Drushku, co-supervised with Nicolas Labroche and Patrick Marcel.
The proposal was published at CAISE [Drushku et al., 2017] and extended at Information Systems [Drushku et al., 2019].

In this section we present an approach for learning users' intentions in a query workload and leveraging them for query recommendation. As presented in previous section, users' intentions come under the *relevance* aspect of interestingness. Therefore, our proposal concerns the discovery of users' interests from the prism of relevance.

We learn a similarity measure intended to capture whether two queries reflect a same user interest, and pair it with a off-the-shelf clustering algorithm. We then use the learned interests for recommending interesting queries.

Considered workloads. We experiment on the Enterprise workload, described in Appendix A, which logs queries of beginners and expert users answering to 10 business needs (named Q_1 to Q_{10}), each corresponding to a specific user interest.

In the context of user interest discovery, the business needs Q_1 to Q_{10} serve as our ground truth, our objective being to cluster together queries (potentially from different user sessions) that addressed the same business need. To be realistic, business needs were defined to expect some overlap in terms of accessed data and queries.

Next subsection introduces a representation of BI queries and interactions, and the following subsections describe our proposal for interests discovery and recommendation.

3.3.1 Representation of BI interactions

This subsection presents our model of BI interaction. Given the proximity of BI interactions in modern BI systems and web searches, our modeling of BI interactions is inspired by the modeling of web search sessions [Guha et al., 2015].

In the context of BI, we consider that an interaction relies on a sequence of keyword queries over some data sources. Each keyword query produces an ordered set of formal queries suggested from the set of keywords. One of these formal queries, chosen by the user, is evaluated over the data source, and then, the answer retrieved is displayed to the user. All this (keyword query, suggestions and chosen query) is called an observation. We extract a set of features that describe each observation of all users' interactions.

Let D be a database schema (resulting from the integration of several data sources), I an instance of D and Q the set of formal queries one can express over D. We call a *database entity* to either an attribute of D or a constant appearing in I. The result (or answer) of a query q over a database instance I is denoted by $q(I)$.

A *BI question* (or question for short), K, is a set of tokens (or keywords) entered by a user. Each token may be matched with database entities to generate queries. To simplify, we describe a multidimensional query q as a set of query parts, as in [Aligon et al., 2011]. A *query part* is either a level of a hierarchy used for grouping, a measure, or a simple Boolean predicate used for filtering. In what follows, queries are confounded with their sets of query parts, unless otherwise stated.

Example 3.1 *Starting from the question "Revenue for France as Country" the following tokens* $K_1 = \{$*"Revenue", "for", "France", "as", "Country"*$\}$ *are identified. A corresponding formal query contains the following query parts: Revenue is a measure, Country a level in a hierarchy, and* $France$ *is a constant, resulting in Country=France being a Boolean predicate.* □

As tokens are entered, a BI system might on the fly suggest further tokens to complete the current ones, letting the user choose among them, as in web search engines. The underlying idea is that a suggestion completes the original BI question to obtain a well-formed query over a database.

We formalize the notion of suggestion as follows.

Definition 3.2 (Suggestion) *A* ***suggestion*** *is a triple* $s = \langle K, D, q \rangle$ *where* K *is a BI question,* D *is a database schema and* q *is a query over* D*.* □

Example 3.2 *The question* $K_1 = \{$*"Revenue", "for", "France", "as", "Country"*$\}$ *is completed to focus on year 2017. The corresponding suggestion,* $s_{11} = \langle K, D, q \rangle$*, consists of question* $K = \{$*"Revenue", "for", "France", "as", "Country", "and", "2017", "as", "Year"*$\}$*, schema* D*, which includes a relation Sales, and the formal query* q*, represented by its three query parts {Revenue,* $Country = France$*,* $Year = 2017$*}, whose SQL code is:*
SELECT sum(*Revenue*) FROM Sales WHERE *Country='France*' AND Year=*2017*; □

In Web Search, information needs are modeled as sequences of observations, an observation being a search engine query with its associated web results (or SERP) and clicks [Guha et al., 2015]. We adapt such model to BI interactions. This adaptation relies on the following simple analogy: (i) the search engine query corresponds to the BI question, (ii) the SERP corresponds to the set of suggestions associated with the BI question, and (iii) a click on one SERP link corresponds to the choice of a suggestion and hence to the evaluation of the query associated with the suggestion.

Formally, we define observations and iteractions as follows:

Definition 3.3 (Observation) *An* ***observation*** *is a triple* $o = \langle K, S, s \rangle$ *where* K *is a BI question,* $S = \{s_1, \ldots, s_n\}$ *is a set of suggestions for question* K*, and* $s \in \{s_1, \ldots, s_n\}$ *is the suggestion selected by the user.* □

Definition 3.4 (Interaction) *An* ***interaction*** *of length* v *is a sequence of* v *observations* $i = \langle o_1, \ldots, o_v \rangle$ *that represents the user interaction with the BI system.* □

Remark that an interaction is a particular kind of user session, where not only database queries are logged, but the whole observations.

Finally, we define a user interest as follows:

Definition 3.5 (User interest) *A* ***user interest*** *is a finite set* $U = \{o_1, \ldots, o_n\}$ *of observations that represents one particular information need.* □

These concepts are at the core of our proposal for clustering observations, as explained in next subsection.

3.3.2 Clustering observations

We formalize the problem of discovering coherent user interests as a clustering problem, for which a similarity measure is learned over a set of descriptive features. These features allow observations (and user interests) to be grouped based not only on the expression of the BI question but also based on their intent as expressed by the chosen suggestion, and on their knowledge, as provided by the evaluation of the chosen query. To compare two user interests, a global similarity is computed as a weighted sum of feature-based similarity measures.

In this subsection we first define the set of features we consider, and then we address two main problems: (i) determining a similarity measure between user interests and, (ii) finding a clustering algorithm that can work on the sole basis of this similarity.

User interest features. To provide the best characterization of user interests, we define a set of candidate features, that we subsequently analyze to identify those maximizing the accuracy from the user's perspective. Features should be understood as dimensions on which it is possible to compare two user interests, $U_1 = \{o_1^1, \ldots, o_l^1\}$ and $U_2 = \{o_1^2, \ldots, o_m^2\}$.

We considered three groups of features, listed in Table 3.3. The first group relates to the BI questions and suggestions (Features 1-6). The second group relates to the chosen suggestions, and especially their query parts (Features 7-9). Both groups proved effective in identifying interests in Web searches [Guha et al., 2015]. The third group consists of specific BI features and relates to formal queries and their answers (Features 10-15). For each feature, we propose a similarity measure that is the most suited for it (e.g., cosine for vectors of frequencies, Jaccard for sets).

#	**Feature**
1	Frequency of tokens
2	Frequency of refining tokens
3	Suggestions
4	BI questions
5	U_1 questions that are sub-questions in U_2
6	U_1 questions in the same interaction as a question in U_2
7	Frequency of chosen query parts
8	Frequency of tokens of U_1 that match chosen query parts of U_2
9	Chosen suggestions
10	Levels in chosen query parts
11	Tuples retrieved by chosen queries
12	Queries in U_1 that differ by one query part from a query in U_2
13	Sources
14	Attributes of U_1 functionally identifying attributes in U_2
15	Expertise of users

Table 3.3: Features considered

Learning observation similarity. For the first problem, i.e. determining a similarity measure, our aim is to distinguish among the candidate features presented above, those who are the most suitable to identify coherent interests from a user's standpoint. As we are expecting an explainable model that provides the relative importance of each feature, we rely on a linear aggregation for our similarity $Sim(U_1, U_2)$ defined as follows:

$$Sim(U_1, U_2) = \sum_{f=1}^{n} \omega_f v_f(U_1, U_2) \tag{3.1}$$

where n is the number of features, v_f is the similarity measure for feature f and ω_f is a weight representing this feature's importance in the comparison.

With this formulation, the problem of designing a similarity naturally translates into a problem of determining the set of weights ω_f paired with each similarity measure v_f.

To this end, we formalize the problem of discovering ω_f as a classification task, which proved effective in [Guha et al., 2015, Wang et al., 2013]. Indeed, we are able to train a classifier (X, Y) in which each entry $x \in X$ corresponds to a couple of observations, the descriptive features of each entry being the ones introduced in Table 3.3 and the output $y \in Y$ being set to 1 if these two observations relate to the same interest, and -1 otherwise.

We use an off-the-shelve SVM linear classifier paired with some ground truth knowledge about users' interests to learn the predictive value of the features. For a feature f, the weight ω_f is set to the conditional probability that two observations correspond to the same user interest knowing that they coincide on feature f. The absolute value of ω_f reflects how discriminant feature f is (a large value indicates that feature f is very influential in the decision process), while the sign of ω_f denotes that feature f will either act in favor of grouping user interests or, conversely, to separate them. In particular, the descending list of the absolute value of weights ranks the features, stating from the most important one.

Clustering. The clustering problem is addressed by experimenting with off-the-shelf well-known and trusted relational clustering algorithms implementing different strategies, i.e., centroid-based, connectivity-based and density-based clustering.

3.3.3 Recommendation of queries

To illustrate the practical use of our approach, in this section, we present *IbR* (Interest-based Recommender), a simple recommender specifically designed to exploit the clusters that represent user interests. IbR is inspired by and adapts previous approaches proposed to predict or recommend OLAP queries.

First, inspired by Falseto, a collaborative recommender described in [Aligon et al., 2015], IbR recommends a sequence of queries representing the sequence of moves that is expected to best complete the beginning of an interaction. As remarked in [Aligon et al., 2015], it is expected that users, especially non-expert ones, benefit from a sequence of recommended queries, in that it gives them a compound and synergic view of a phenomenon, carries more information than a single query or set of queries by modeling the potential expert user's behavior after seeing the result of the former query.

Second, we borrowed from the work of Sapia [Sapia, 2000] and the work of Aufaure et al. [Aufaure et al., 2013b] the idea of using an order-1 Markov model to probabilistically represent user behaviors. Like in the latter [Aufaure et al., 2013b], the states of the Markov model are clusters constructed from a set of past interactions, with two notable differences: (i) observations are used in our case, instead of queries, and (ii) clusters correspond to coherent interests. IbR can be seen as a model that guides the user's next moves based on the probabilities of moving between discovered user interests.

By construction, we expect IbR to have two types of benefits: sharing expertise between users, and recommending queries that are diversified in terms of interest.

Principle. The principle of IbR follows the same two-phase approach than that of [Aufaure et al., 2013b]. The first phase is off-line and consists of clustering the observations to detect user interests, as detailed in previous section. Then, clusters are treated as states of a Markov chain model, and the probabilities of the most likely next state are computed as explained below.

The only on-line phase of the recommender is when a new interaction begins, each observation of the interaction is used to compute the most likely query in the sense of the Markov model. For the rest of observations, a comparison is not necessary since the recommendation only derives from the previous recommendations, i.e., the last state calculated by the Markov model.

Learning the Markov model. The creation of the Markov model is done as follows. Let U be the set of clusters expressing user interests. The states of the Markov model are the clusters of U. The transition probability distribution is given by $\Pr(X_{n+1} = x \mid X_n = y) = \frac{n_{xy}}{n_y}$ where x and y are clusters in U, n_y is the size (the number of observations) of cluster y and n_{xy} is the number of interactions that contain two adjacent observations o_i, o_{i+1} such that o_i is in cluster y and o_{i+1} is in cluster x. We use a special state to represent the end of interactions, which is used to obtain the probability of ending the recommendation.

The prediction algorithm. Given an observation, called the current observation (whose chosen query is called the current query) from now on, we identify the user interest (i.e., the cluster) that this observation is the closest to by computing the average similarity between the current observation and all the observations of each cluster.

Once we have identified the cluster, the Markov model gives the most likely next state. By construction, since states coincide with user interests, it is expected that the most likely next state is the current one. To distinguish between the two types of benefits our recommender can have, we devised two strategies for generating the recommended sequence, reflected in the two modes our recommender can operate. Mode 1, named *IbR1*, tries to benefit from the expertise coming from this next probable cluster only. Conversely, Mode 2, named *IbR2*, tries to anticipate when users change their focus (e.g. to address other business questions) and propose recommendations diversified in terms of user interests. We now describe these two modes precisely.

IbR1 forces the recommender to choose the queries for the recommended observations in the next probable state only. In other words, IbR1 does not use the Markov model but uses only interest identification. The chosen queries are ordered by decreasing similarity to the current query. The length of the recommended sequence is ruled by a similarity threshold that ends the sequence if the similarity between two consecutive queries is considered too small.

The second mode, IbR2, fully uses the Markov model based on user interests. In other words, IbR2 acknowledges the fact that user interactions may span across different interests and composes the recommended sequence of queries as follows:

1. the first query of the sequence is the chosen query of the observation that is the most similar to the current observation;
2. this observation is used as the new current observation for which the next interest is identified with a random draw using the Markov model, which means that the probability to reach another interest is low, not null, which is different from IbR1;
3. the most similar observation of the next probable state, according to the Markov Model, among those not yet recommended, is identified and its chosen query is added to the sequence;
4. this algorithm iterates until the final state of the Markov model is reached.

Metrics for assesing recommendation quality. To evaluate our approach, we rely on the literature on recommender systems [Herlocker et al., 2004, Baeza-Yates and Ribeiro-Neto, 2011, Gunawardana and Shani, 2015] as well as on a protocol specially conceived for comparing recommendations of query sequences [Aligon et al., 2015]. We measure two of the most

commonly employed criteria to judge the recommendation quality to assess whether our recommender is able to achieve a good balance between the ability to recommend and the quality of its recommendations, namely:

- accuracy, i.e., the degree to which recommendations correspond to what is expected in terms of queries, and
- coverage, i.e., the degree to which recommendations can indeed be generated.

We enrich this set of measures with the following criteria to understand whether our recommender favors expertise sharing between users and interest diversity:

- expected diversity, i.e., the degree to which recommendations correspond to what is expected in terms of user interests,
- expected user, i.e., the degree to which the current user is retrieved in the recommendations,
- expertise, i.e., the degree to which recommendations come from experts, and
- expertise benefit, i.e., the degree to which beginners can benefit from expert recommendations.

Regarding accuracy, acknowledging that finding the exact next query of an interaction is very unlikely, we use extended versions of Precision and Recall measures to incorporate similarity between interactions and queries. Given an interaction i, let f_i be its actual future (i.e., the sequence of queries the user would have formulated after the last query of i if they had not been given any recommendation) and r_i be a recommended future. Recommendation r_i is considered to be *correct* when $r_i \sim f_i$, i.e., when it is similar to the actual future of i.

We use the similarity measure proposed in [Aligon et al., 2014b][3] as it is independent from our proposal and can fit any recommender system under testing, contrary to ours, which needs proper interactions to work.

Details on the computation of quality metrics can be found in [Drushku et al., 2019].

3.3.4 Experiments and results

In this section we report the major findings of our experiments for clustering observations and recommending next queries.

In what follows, we consider the Enterprise workload. The complete workload, dubbed Complete hereafter, contains 24 user interactions, each one possibly chaining several business needs. It accounts for 530 observations. To have several difficulty settings, we also built two reduced workloads named Reduced1 and Reduced2, each corresponding to 4 distinct business needs and 4 distinct data sources, which in turn removes most of the potential overlap. Each of them contains 225 observations.

Protocol. Our first objective is to determine a similarity measure based on the features introduced in Table 3.3 that allows, when paired with a clustering algorithm, the grouping of user observations into clusters that accurately reflect user interests. Our second objective is to use these clusters for recommending queries that share the same user interests.

In this regard, our first experiment aims to determine and validate the best subset of features from the set presented in Table 3.3, both, in order to avoid any problem of overfitting when the number of dimensions increases, while still maximizing the quality of the discovery of user interests. To this aim, we test several subsets of features and train the weights of the similarity measure with a linear SVM algorithm, as presented in Subsection 3.3.2, on the sole basis of these

[3] We already used this measure in Section 2.5 to compare with our measure for learning analysis patterns.

features. We also compare several clustering algorithms in order to choose the one that best performs when paired with the learned measure. As no hypothesis can a priori be made on the shape of expected clusters of observations, we use in our tests various clustering algorithms that are representative of the diversity of common methods from the literature. For the comparison, we use Precision, Recall and ARI measures.

We also study how our method handles previously unseen business needs and how general the learned measure is. To this aim, we consider Reduced1 and Reduced2 workloads (which cover different business needs, with few overlap among them), using one workload to train the measure and the other to test the clustering.

Then, we perform a comparative experiment with the state-of-the-art similarity measure for OLAP sessions proposed in [Aligon et al., 2014b], dubbed AD as in Chapter 2.

Finally, we take advantage of the identified user interests to recommend a sequence of observations to users, to help them continue their explorations. We experiment the two recommendation modes introduced in Section 3.3.3, IbR1 and IbR2, and we compare them with two state-of-the-art query recommender systems, QueRIE [Eirinaki et al., 2014] and Falseto [Aligon et al., 2015], both based on collaborative filtering and kNN. As QueRIE recommends a set (instead of a sequence) of queries, we rank the recommended queries according to their similarity to the current session and arrange them in a sequence as was done for IbR1. Note that this transformation of QueRIE output is necessary to ensure that it is comparable to the other recommenders and is under the same conditions. Conversely, Falseto directly recommends a sequence of queries. We notice that unlike IbR1, IbR2 and QueRIE, Falseto does not directly recommend queries that are simply picked in a log file of past queries. Indeed, it picks queries from a log file and then modifies these queries to align them with the current interaction. We expect Falseto to explore more globally the space of possible queries as it builds new queries (not necessarily existing in the log) based on current queries.

Implementation and setting. Our approach is implemented in Java. We also use Python Scikit Learn [Pedregosa et al., 2011] linear SVM to learn the weights of our similarity measure and R clustering packages `cluster` and `fpc`.

The feature weights are learned over 50% of observations chosen randomly, with a balance in the number of observations per business needs. Additional preprocessing and optimizations are performed to ensure that the SVM is accurate. First, our interests pairs are balanced to guarantee that there are the same number of couples related to the same user interest (labeled 1) as the couples related to different user interests (labeled -1). Second, the hyper parameter C is optimised by an extensive cross validated random search.

Our recommender is built as a Markov model over the interactions whose observations have been clustered to identify user interests. Consistent with the protocol proposed in [Aufaure et al., 2013b], we remove from the set of interactions the ones consisting of only one observation.

Recommendation is tested using a leave-one-out cross-validation approach, as follows: We iterate over a set I of interactions by (i) picking one interaction $i \in I$; (ii) taking one of its prefix i_n of size n as one current interaction and the remaining subsequence f_i as one actual future, with $n \in \{1, |i|\}$; (iii) finding a recommendation r_i for i_n using the remaining interactions, $L \setminus \{i\}$. If such a r_i exists, it is compared to f_i computing $r_i \sim f_i$.

In our tests, the similarity between interactions is parametrized by a threshold varying in [0,0.9]. This threshold controls the extent to which two interactions should be considered similar.

We use our own implementation of QueRIE and Falseto, tuning their parameters in order to achieve the best accuracy.

Measure learning and choice of clustering algorithm. Our experiments show that a particular subset of features (henceforth dubbed *G2*), principally composed of features related to business objects and formalized queries (listed in Table 3.4), in collaboration with PAM clustering algorithm [Kaufmann and Rousseeuw, 1987], better identifies the user interests. Adding more features in the observations comparison reduces the accuracy of the measured similarity between them, which leads to interests that are non well-defined. The detailed comparison of several subsets of features and clustering algorithms can be found in [Drushku et al., 2019].

G2 features include the entered tokens (Feature 1) and the suggestions proposed (Feature 3), but the real difference between user observations is specified by the chosen suggestion (Feature 9) with the query parts composing it (Feature 7) and their matching tokens (Feature 8).

	Selected Features	**Weights**
1	Frequency of tokens	0.39
3	Suggestions	0.41
7	Frequency of chosen query Parts	1.23
8	Frequency of tokens ou U_1 that match chosen query parts of U_2	0.38
9	Chosen suggestions	0.40

Table 3.4: Weights for the best subset of features

Clustering quality. Table 3.5 (first line, concerning G2 features) reports clustering quality in terms of Recall, Precision and ARI, for the Complete and Reduced1 workloads.

As expected, we obtain lower results for the Complete workload, specially in terms of Recall. Indeed, we know from the protocol that business needs heavily overlap. Thus, our method, based on SVM, cannot find a proper linear separation between observations related to different user interests. In contrast, within the Reduced1 workload, observations of business needs are clearly separable, and the problem is much easier for the linear SVM.

In addition, intra clusters dissimilarities (not reported here) are lower than inter clusters dissimilarities. This result is confirmed by the Silhouette coefficient [Rousseeuw, 1987], presented in Table 3.6, which is positive for all the clusters, verifying that the built clusters are cohesive and the majority of observations in each of them are well classified in their own clusters. Regarding the diameter, it is well balanced among the discovered clusters, with some minor differences being explained by outliers.

The results of the generalization experiment (for handling unseen business needs) are shown in Table 3.7, evidencing that our measure is indeed general and can adapt to new business needs as there is no drop in performance between each of the generalization tests. Moreover, the results are comparable to those observed in previous tests.

	Complete 10 clusters			**Reduced1** 4 clusters		
Features	Recall	Precision	ARI	Recall	Precision	ARI
G2	0.51	**0.50**	**0.44**	0.70	0.64	0.54
ALL	**0.52**	0.46	0.42	0.73	0.64	0.56
AD	0.39	0.20	0.14	0.41	0.33	0.10
G2 + AD	0.45	0.43	0.38	0.69	0.62	0.52
ALL + AD	0.40	0.40	0.32	**0.78**	**0.65**	**0.63**

Table 3.5: Comparison of our measure based on G2 features with other measures when paired with PAM clustering. ALL denotes the set of all 15 features, AD is the state-of-art measure [Aligon et al., 2014b] and "+" indicates a measure with added features and corresponding weights.

	Cluster IDs									
	1	2	3	4	5	6	7	8	9	10
#observations	48	77	53	74	58	60	78	34	30	18
Silhouette	0.09	0.08	0.01	0.14	0.09	0.08	0.1	0.13	0.15	0.27
Diameter	0.71	0.72	0.86	0.76	0.76	0.78	0.83	0.70	0.69	0.63

Table 3.6: Silhouette coefficients and diameter for 10 clusters obtained with PAM and the learned measure.

Training	**Testing**	**Recall**	**Precision**	**ARI**
Reduced1	Reduced2	0.73	0.71	0.62
Reduced2	Reduced1	0.76	0.67	0.61

Table 3.7: Generalization of our approach. Each test corresponds to the training of the measure and discovery of user interests on different subsets of business needs.

Further experiments are discussed in [Drushku et al., 2019], in particular, we investigate the behavior of our similarity measure for detecting intra-interaction interests. As expected, results show that increasing the number of clusters, precision increases, while recall and ARI decrease. But interestingly, the composition of clusters in terms of users with different expertise remains very acceptable (e.g. when precision reaches 95%, more than 63% of clusters have users with different expertise). In other words, this shows that our measure can be used to identify shared sub-tasks (or intra-interaction interests), where some experts' queries could be recommended to beginner users having to solve the same business need.

Comparison to a reference similarity measure. Table 3.5 shows how our metric compares to AD measure, designed for comparing OLAP sessions [Aligon et al., 2014b].

We observe 2 distinct behaviors depending on whether we consider the Complete or Reduced1 (where business needs are well separated) workloads. With the Complete workload, our measure with G2 features performs better than the other measures, as it only relies on the most discriminating features. In this particular context of heavy overlap of user interests, adding more features makes the problem even more complex to solve for the linear SVM. In contrast, with the Reduced1 workload, as user interests are clearly separable, the problem is much easier for the linear SVM and adding features may help in finding a better solution by fine tuning the separation hyper plane. Consequently, in this case, slightly better results may be achieved with features other than those of G2.

Incidentally, our experiments also reveal that considering AD measure as a feature in our similarity measure in some cases improves the overall quality of our approach.

However, we expect our approach to be the most efficient in any scenario as the hypothesis that clusters of observations are clearly separated is too strong in practice. Thus, the measure based on G2 features seems to be the most appropriate among those that we evaluated, in particular when compared to the state-of-the-art AD measure [Aligon et al., 2014b].

Recommendation quality. Table 3.8 presents the Markov model. Specifically, it shows the transition probabilities between clusters (states), sources in rows and targets in columns. Note that, as this model is recreated several times in our tests, we present here the model learned over the whole log. It is easily perceived that, as expected, observations of a cluster are mainly followed by observations of the same cluster, meaning that interactions tend to remain within the same user interest.

Figure 3.5 reports the measures of the various criteria defined to assess the quality of the recommenders. We start by discussing these measures for IbR1 and IbR2.

	State 1	State 2	State 3	State 4	State 5	State 6	State 7	State 8	State 9	State 10	Final State
State 1	**0.40**	0.13	0.06	0.0	0.08	0.13	0.02	0.04	0.0	0.0	0.14
State 2	0.10	**0.58**	0.06	0.04	0.05	0.04	0.03	0.03	0.01	0.04	0.02
State 3	0.0	0.06	**0.43**	0.02	0.09	0.13	0.06	0.0	0.13	0.04	0.04
State 4	0.0	0.03	0.05	**0.68**	0.05	0.03	0.05	0.0	0.05	0.02	0.04
State 5	0.02	0.07	0.02	0.05	**0.54**	0.03	0.15	0.10	0.02	0.0	0.0
State 6	0.02	0.17	0.02	0.08	0.07	**0.46**	0.15	0.03	0.0	0.0	0.0
State 7	0.04	0.0	0.10	0.14	0.0	0.1	**0.55**	0.03	0.0	0.0	0.03
State 8	0.03	0.03	0.08	0.0	0.06	0.03	0.15	**0.50**	0.06	0.0	0.06
State 9	0.17	0.03	0.03	0.0	0.03	0.07	0.03	0.10	**0.50**	0.0	0.04
State 10	0.11	0.06	0.11	0.0	0.0	0.06	0.0	0.0	0.0	**0.39**	0.27

Table 3.8: Transition probabilities for the 10 states (clusters) of the recommender's Markov model

The coverage is as expected. By design, IbR1 achieves a perfect coverage, while IbR2, using a probability for ending the session, may not recommend, particularly for a longer current session. Regarding accuracy, both recommenders perform very well, with IbR1 performing the best, when the similarity threshold is set low (0.4 or below). Below this threshold, both recommenders show the same behavior.

In terms of expected diversity, as expected, IbR2 outperforms IbR1 since the latter cannot move outside a current interest. Notably, even for quite demanding similarity thresholds, IbR2 performs reasonably well in predicting interest switches.

The very low scores for both recommenders in terms of the expected user is expected in that it confirms that none of them were designed to stick to the current user. Nevertheless, we note that both IbR1 and IbR2 still do better than state-of-the-art recommenders for low similarity thresholds, which can be interpreted as a side effect of user interest detection.

Both recommenders perform well in terms of expertise, with IbR2 being more robust than IbR1 to the similarity threshold. This is due to IbR2 being more likely to find expert queries in clusters other than the current one. Finally, both recommenders perform fairly in recommending expert queries to beginners. We note that they are not designed to do so and good performances for this criterion would be a side effect of clustering interests. However, extending the recommenders to favor this behavior can be done easily if expertise is recorded or can be deduced from the observations.

In summary, IbR1 performs slightly better in terms of accuracy and coverage, while IbR2, with its global exploration of the user interests, is better at identifying interest switching and proposing recommendations coming from expert users.

Comparison with state-of-the-art recommenders. We now discuss how IbR1 and IbR2 compare to two state-of-the-art recommenders.

Comparison with QueRIE. QueRIE achieves perfect coverage, as does IbR1, and it is similar to it in terms of accuracy for low similarity thresholds, being slightly more robust to more demanding thresholds. This similarity of behavior can be explained by the nature of both recommenders, which are very similar as they tend to locally explore the user interest based on current queries.

We note that QueRIE is always better than IbR1 for expected diversity, since the latter is bound to a specific interest, and is slightly better than IbR2 for very low similarity thresholds but is expectedly less robust than it to high thresholds. Finally, as expected, its results in terms of expected user, expertise or expertise benefit show that it has not been specifically designed to take these features into account.

Figure 3.5: Coverage, accuracy, expected diversity, expected user, expertise and expertise benefit for IbR1, IbR2, Falseto and QueRIE.

Comparison with Falseto. Among all the recommenders, Falseto achieves the worst performances both in terms of coverage and accuracy for low similarity thresholds. Regarding coverage, it is clearly impacted by the demanding session similarity measure that Falseto internally uses to align current and past sessions to generate candidate recommendations. Indeed, when query similarity is below Falseto's built-in threshold, no past session is found to be similar to the current one, which results in no candidate recommendations, which disables the recommendation.

Remarkably, Falseto is more robust in terms of accuracy when the similarity threshold becomes more demanding. This can be explained by its fitting phase, which aligns the recommendations with the current interaction, i.e., even if the candidate recommendation picked from the log is not the one expected, the fitting phase is able to sufficiently modify it to bring it closer to the expected future. As expected, Falseto is outperformed in terms of expected diversity, expected user, and expertise, but surprisingly achieves the best expertise benefit. This can be because its candidate recommendations are sequences that are similar to others in the log and that such sequences are more likely produced by expert users.

Further experiments are discussed in [Drushku et al., 2019], in particular, we investigate whether leveraging user interests calls for a tailored recommendation strategy or can benefit an existing one. To this end, we give Falseto and QueRIE the chance of knowing the user interests beforehand. More precisely, we force them to recommend queries inside each cluster separately and to simulate their behavior if they were not agnostic of user interests. The results show that detecting user interests may be useful for already existing recommendation strategies. This is particularly clear for QueRIE, which performs better in the majority of cases when the interest is leveraged. This is more contrasted for Falseto, which compares to whole sessions in the log. Therefore, when only one cluster is available, Falseto will misses those sessions spanning different clusters.

3.3.5 Discussion

This section presented a collaborative recommendation approach that leverages users' interests in modern BI systems to relieve the user from tedious explorations. This system combines state-of-the-art techniques from literature in Web Search and BI query recommendation. At the heart of it is an approach for identifying coherent interests of BI users with various expertise querying data sources by means of keyword-based analytical queries. Our approach relies on the identification of discriminative features for characterizing BI interactions and on the learning of a similarity measure based on these features. Once user interests are identified, they are treated as first-class citizens in a collaborative BI query recommender, that suggest next moves in an exploration based on the probability for a user to switch from one interest to another.

We have shown through user tests that our approach is effective in practice and can be beneficial to analysts whose interests match those of expert users, or whose interests change during the analysis. Overall, our results show that keyword-based interaction systems provide semantically rich user traces well adapted to the detection of coherent BI users' interest and that such interests can also be exploited successfully by state-of-the-art recommendation strategies.

Building upon these results, many practical benefits of our approach can be envisioned, going beyond keyword-based interaction systems. In particular, we envision the design of an intelligent assistant that raises alerts when data sources are refreshed or when users' information needs and expertise change. This work can be an inital step for investigating new interest- and skill-based recommendation approaches.

Finally, we observe that even if the approach proposed in this section is based in the discovery of users' intentions (pertaining to relevance, a particular aspect of interestingness), it can be adapted to consider or combine other interestingness aspects, exploiting other user models (e.g. users' belief) in addition to the query history. New features are surely necessary, and adapted similarity measures must be therefore trained. But after user interests are discovered, the Markov model can be learned in the same way, and IbR1 and IbR2 can be directly applied on such interests, or at least they can serve as a baseline for benchmarking new recommenders.

3.4 Conclusion

This chapter presented our contributions for understanding, modeling and learning users' interests.

We first presented our findings on an extensive survey of the literature, both in the area of computer science and in the area of the study of human behavior. We proposed a two-level framework for developping interestingness measures. At the first level, we proposed 4 high-level aspects of interest, and at the second level, we developed several data-oriented assessment algorithms, showing how even simple measures can help the analysis of users' interests.

We then focused in a particular aspect, relevance, and we proposed an approach for learning users' interests in a query workload and exploiting them in a query recommender. We used classification, clustering and recommendation techniques, which succeded to capture users' intentions, being effective in practice, and specially beneficial to novice analysts.

Table 3.9 summarizes our contributions in terms of query model, used workloads and main lessons learned.

Contribution	**Proposed models**	**Query language**	**Used workloads**	**Main lessons learned**
Interest model	Aspects	—	Open, Adult, Loan	Correlation with focus and exploration quality. No strong aspect preference. Perceptions change over time.
	Measures	OLAP-like		
Interest learning	Clustering	OLAP-like	Enterprise	Similarity measure based on suggestion-related features. Outperforms state-of-the-art measure.
	Recommender			Good quality. 2 well-identified modes. Outperforms state-of-the-art recommenders.

Table 3.9: Summary of contributions

Our proposal for recommending queries based on users' interests was refined in industrial context and patented [Drushku et al., 2021].

In addition to this direct application, our framework and techniques have been used for recommendation in other applications contexts. In particular, we can mention:

- The Intentional Analytics Model (IAM) [Vassiliadis et al., 2019] has been recently envisioned as a new paradigm to couple OLAP and analytics. One of the pillars of IAM is returning enhanced query results, i.e., multidimensional data annotated with knowledge in the form of interesting model components (e.g., clusters). In [Francia et al., 2022c], we developed a proof-of-concept for the IAM vision by delivering an end-to-end implementation of *describe*, one of the five intentional operators introduced by IAM.
  The interest of a component is computed as a weighted sum of novelty, peculiarity and surprise measures, tailored for IAM context. The most interesting components are highlighted.
- The Traveling Analyst Problem (TAP) [Chanson et al., 2020], is an original strongly NP-hard problem where an automated algorithm assists an analyst to explore a dataset, by suggesting the most interesting and coherent set of queries that are estimated to be completed under a time constraint. Similarly to automated machine learning, TAP aims at (i) finding, from a very large set of candidate queries, a subset of queries that maximizes

their interest within a limited time budget, and (ii) ordering them so that they narrate a coherent data story.
A crucial part of TAP lies in the definition of an interestingness measure to determine the optimal subset of queries. Such measure must be quickly computed before the actual evaluation of the queries, and therefore it relies on the text of the query. We use an innovative measure of surprise, based on prior knowledge beliefs on query parts.

– In [Chanson et al., 2022b], we proposed an approach for generating personalized data narrations by extracting messages from a collection of EDA notebooks over a given dataset. The approach consists of extracting features from notebooks to learn what interesting messages they expose and then producing a user-tailored data narration, i.e., a coherent sequence of messages matching a given user profile.
An interestingness model was learned from notebook and messages features (such as notebook popularity and structure, and measure complexity and explainability), using regression models and auto-machine learning.

Our interest-based recommendation approach has been studied in several domains, including: customer segmentation [Carbajal, 2021], profiling of users' beliefs [Chanson et al., 2019], conversational, self-service, intentional and emotional BI [Pinon et al., 2022, Francia et al., 2022a, Vassiliadis et al., 2019, Bimonte et al., 2023], volunteer data warehousing [Sakka et al., 2021a], data lake exploration [Gunklach et al., 2023b]. It has inspired the recommendation of points of interest [Gan and Ma, 2023], chain composite items [Chanson et al., 2021], and business data [Pinon, 2023], and used for natural language BI recommendation [Guessoum et al., 2022].

Next chapter goes beyond EDA support, and investigates the overall process for narrating data stories. Users' interests, and in particular users' intentions, are at the kernel of our proposal.

Chapter 4

Modeling data narratives

This chapter describes our contributions for understanding and modeling data narratives.

It relies on materials published in several conferences and journals, the main ones being [Outa et al., 2020b, Outa et al., 2023]. The overall contributions were developed in collaboration with several PhD and master students, as well as researchers and journalists, as summarized below.

Advising, projects and collaborations

PhD theses:
Faten El Outa (2019-2023), *A framework for crafting data narratives*, co-supervised with Patrick Marcel.
Raymond Ondzique Mbenga (2019-2023), *Business Intelligence system, from narration to simulation: Application to epidemic surveillance of Tuberculosis in Gabon*[1], co-supervised with Thomas Devogele and Edgar Brice Ngoungou.

Master theses and projects: Lucile Jacquemart (2021), Bassem Salloum (2021), Valentin Fradet (2022), Jimmy Rata Gobal (2023).

Research projects:
Madona – *Madona – Mastering Interactive Data Analysis for Journalistic Narration*[2] (2018-2022), national funding (MaDICS, CNRS).

Main collaborations:
Panos Vassiliadis (University of Ioannina, Greece),
Matteo Francia (University of Bologna, Italy),
Edgar Brice Ngoungou (University of Health Sciences, Gabon)
Marie Chaignoux (University of Paris, France),
Raphaël Da Silva (Rue89Strasbourg Newspaper, France).

[1] Written in French. Original title: *Système d'information décisionnel, de la narration à la simulation : application à surveillance épidémiologique de la tuberculose au Gabon*

[2] Original name (in French): *Maîtriser l'Analyse interactive de Données pour la Narration journalistique*
`https://sites.google.com/view/action-madics-madona/home`

Contents

4.1 Problems and positioning

Narrating a story is considered as one of the oldest activities in the world, and a pillar of information communication as a mean of education. Often mistaken with storytelling, which describes the social and cultural activity of sharing stories[3], narration is the use of techniques to convey a story to an audience[4].

More recently, data narration, i.e., narrating with data visualizations [Hullman et al., 2013], received increasing interest in several communities (e.g. Journalism, Business, e-Government, Health). Data narratives, i.e. the outcome of data narration, are largely used by journalists, scientists, and other communicators, to convey striking messages to a given audience. They may take the form of a data video, an infographics, a news article, etc., and more generally, any sort of narrative that is crafted based on data can be considered a data narrative.

While using many terms (e.g., visual data narration, narrative visualization, visual storytelling, data driven storytelling), the data visualization community has brought much attention to data narration [Carpendale et al., 2016]. Very recently, data science and machine learning communities interested to the topic, mainly under automation lenses [De Bie et al., 2022]. We claim that EDA techniques should be at the kernel of data narration support tools.

Actually, data narration includes a variety of activities, including the analysis of data, the drawing of relevant messages from data, the structuring of messages into a coherent story and its visual rendering. But despite this diversity of activities, sometimes even conducted by different people with varied professions and skills, there is no framework, model, workflow, or tool for holistically supporting the crafting of data narratives. A more global approach to data narration is needed from domains including data visualization, data management, data exploration and machine learning. A very recent survey on data stories also points the need for integrated, cross-disciplinary approaches, and asks for considering cognitive, emotional and contextual impacts [Schröder et al., 2023].

The scope of such an integrated framework targets the population of data journalists or any other data enthusiast that craft data narratives out of existing data. It should provide methodological guidance, enable tool support and recommend actions to less-experienced data narrators. In particular, an application that would automatically document the data exploration and narration crafting is desperately needed by data workers, who spend hours to document their work. This is important for reproducibility, transparency, and linkage, and requires consensual models.

Our research challenge is to **model the static and dynamic aspects of data narration**, setting the bases for the development of data narration frameworks and tools.

Several research needs arise. They are described in the following subsections.

4.1.1 Need for conceptualization of data narrative

Data narration refers to the activity of producing narratives supported by facts extracted from data analysis, using interactive visualizations [Carpendale et al., 2016]. More concretely, such data narratives can be viewed as ordered sequences of steps, each of which can contain words, images, visualizations, audio, video, or any combination thereof, and which are based on data [Kosara and Mackinlay, 2013]. Apart from these general considerations, and to the best of our knowledge, there is no consensual definition of data narrative, let alone a conceptual or logical model of it.

[3] urlhttps://en.wikipedia.org/wiki/Storytelling
[4] `https://en.wikipedia.org/wiki/Narration`

A first problem to investigate is **how to define a data narrative** and then **how to model the key concepts of the domain**.

While various models of narrative have been proposed (see [Elson, 2012] for a survey), none of them qualifies for data narrative. However, some aspects of classical narration theory, as described, e.g. in [Chatman, 1980], should be reviewed to understand the fundamental structure of narrative.

In addition, despite the lack of holistic models for data narrative, a large palette of related concepts have been proposed in the literature (see Section 2.2. of [Outa, 2023] for a review of the state of the art). Even if most of such concepts were described from specific perspectives (e.g. visualization) or concern very specific tasks (e.g. introducing interactivity), they provide a rich base for modeling the domain.

Our goal is to **develop a conceptual model that provides a structured, principled definition of the key concepts of the domain, along with their relationships, and clarifies their role and usage**.

Our research track is to adapt traditional narrative models to match data narratives, and extend them in order to reflect the main concepts proposed in the literature. This model aims to guide a data narrator to craft a data narrative from scratch: fetch and explore data, abstract important messages based on an analysis goal, structure the contents of the data story, and render it in a visual manner. We first study the state of the art in narration, data visualization, data management, data exploration, and computer-human interfaces, among others. Then a tight collaboration with data journalists and communication science practitioners (in the context of the Madona project), including surveys and observation sessions, allows a fine tuning of the proposal.

Section 4.2 presents our contributions for modeling data narratives.

4.1.2 Need for modeling the data narration process

Having the conceptual model in mind, our aim is to study the dynamic aspects of data narration. Like many works in the literature (e.g., [Kosara, 2017, Lee et al., 2015, Chen et al., 2020]), we postulate that the different forms of data narration can be described by a comprehensive process encompassing the various activities ranging from data exploration to the rendering of the data narrative. A formal description of this process will benefit novice data narrators, like e.g., non technical data journalists, and will be instrumental to the development of tools for supporting advanced data narrators.

The problem we investigate is **how to model the data narration process**.

Few works offer comprehensive workflows describing the entire data narration process. The first attempts to model data narration processes come from the data visualization community. For example, [Kosara and Mackinlay, 2013] proposed a two-phase process: First, data narrators collect information and *explore* their interrelationships, pointing to key facts, and then, they *tie* those facts together into a story. Later, [Lee et al., 2015] identified three main phases: *explore data* to retrieve findings, *make a story* to turn findings into a sequence of narrative pieces to build the plot of the narrative, and *tell a story* to materialize the plot in a visual manner. Most works (e.g. [Chen et al., 2020, Wang et al., 2020]) agree on these 3 general phases.

Our bibliographical study revealed the absence of a comprehensive and well-founded process that covers the main activities of the data narration process, specially those dealing with users' intentions and their tight relation to data analysis. However, such intentional activities are very frequent in practitioners processes [Chagnoux, 2020].

Our goal is to **develop a comprehensive and well-founded data narration process, founded on the conceptual model of the domain, that covers the whole data narration cycle and accommodates a wide range of practices observed on the field**.

Our research track is to review the processes and activities described in the literature and confront them to those described by practitioners and observed on the field. As done for developing the conceptual model, a tight collaboration with practitioners (many data journalists within the Madona project, but also data scientists and public health analysts), as well as the observation of students, allow a fine tuning of the proposal.

Section 4.3 presents our contributions for modeling the data narration process.

4.1.3 Scope

We first develop general models, covering a large palette of applications and practices, and keeping independent of particular professions and usages. We then investigate particular scenarios, and study the instantiation of the model to concrete application contexts.

Road map. Sections 4.2 and 4.3 present our contributions to the previously described research needs and Section 4.4 draws our conclusions.

4.2 Conceptual model for data narrative

Foreword

This section summarizes part of the PhD thesis of Faten El Outa, co-supervised with Patrick Marcel, and in collaboration with Panos Vassiliadis and Matteo Francia.
The proposal was published at ER [Outa et al., 2020b], and showcased with a demonstration [Outa et al., 2020a].

This section presents our proposal of conceptual model for data narratives providing, a principled definition of the key concepts of the domain, along with their relationships, and clarifying their role and usage.

It is based on four layers that reflect the transition from raw data to the visual rendering of the data story: factual, intentional, structural and presentational. This model aims to support the entire lifecycle of building a data narrative: fetch and explore data, bring out findings, derive interesting messages, structure the plot of the data narrative, and render it in a visual manner.

Considered narratives. We experiment on Stokes and Covid narratives, described in Appendix B.

Next subsections present the model, including a description of model layers and an example. We use the term *data narrator*, or simply narrator, for referring to the designer of the data narrative, who is not necessarily a business analyst, but can be a data journalist, a data scientist or a plain data enthusiast, aiming to produce a report of findings. We also assume an *audience* for the produced outcome, which includes the people that will see, read or hear the story. Both narrator and audience can represent several persons, or be confounded into one person.

4.2.1 From narrative to data narrative

Narratives. Narrative theoreticians agree that there are at least two levels in any narration: some events happen (what is told) and these events are presented and transmitted to an audience in a certain way (how is it told). They are called respectively story and discourse [Akleman et al., 2015].

Chatman distinguishes narration's elements defining narrative as a couple of a story (content of the narrative) and a discourse (expression of it) [Chatman, 1980]. The story has a form (the story elements, i.e. actions, happenings, characters, settings) and a substance (a composition of story elements as pre-processed by the narrator's cultural code). The discourse has a form (a translation of the story content to a structured combination of the story elements), and a substance (the set of all media used to show structured elements, like text, pictures, tables or charts). In summary, the story can be seen as the logical form of the narrative, while the discourse is its presentable manifestation, obtained through narrator's editions: prunes unimportant parts out, magnifies some others deemed interesting, rearranges the order of presentation to make it more interesting, etc.

Visual data stories. As pointed in [Schröder et al., 2023], the term *storytelling* has been broadly used by the data visualization community without a universally accepted definition, but sharing a common trait of portraying a process or sequence of events. The authors distill the existing definitions and define a data-driven story as a series of related events in a (meaningful) context to facilitate understanding and decision-making concerning data.

[Kosara and Mackinlay, 2013] note that journalists collect information, which gives them the key facts, and then they tie those facts together into a story. The goals, tasks and tools used during the research phase differ from those in the writing phase, and only some of the material from the research phase end up in the final story, most of the source material only serving as raw background information. [Segel and Heer, 2010] insist that the notion of a chain of causally related events is central.

More recently, [Chen et al., 2020] distinguishes (a) visual analytics, which requires to see all aspects of complex data, explore their interrelationships, and is supported by multiple coordinated views and sophisticated interaction techniques, from (b) storytelling, which is meant to convey only interesting and/or important information extracted through the analysis, presented in a simple and easily understandable way. The two processes differ in their purposes, target users, kind of information dealt with, and methods of presenting the information and interacting with it. To support telling stories of visual analytics findings, there should be an intermediate step between analysis and storytelling, in which the narrator assembles and organizes information pieces to be communicated.

Data narratives. Inspired by [Chatman, 1980] and [Chen et al., 2020], we propose the following definition for data narrative:

Definition 4.1 (Data narrative) *A data narrative is a structured composition of messages that (a) convey findings over the data, and, (b) are typically delivered via visual means in order to facilitate their reception by an intended audience.* □

We highlight two important differences with respect to classical narratives: First, data narratives are supported by data, and messages convey findings over data. And second, adapted media (graphics, maps, videos, animations, audios...) are used to convey messages to the audience.

We borrow Chatman's terminology and extend his structure of narrative considering that data narrative must describe how the content of the story (Chatman's elements) is derived from data. This is done by distinguishing 4 layers in our model of data narrative: the first two layers represent the *story* and the last two represent the *discourse*. In the story, a *factual* layer represents the story form while an *intentional* layer represents the story substance. In the discourse, a *structural* layer represents the discourse form and a *presentational* layer represents the discourse substance.

Next subsections respectively describe the proposed model and illustrate its main concepts within an example.

4.2.2 Model description

The proposed model arises from a detailed survey of data narration concepts (reported in [Outa, 2023]) and a tight collaboration with practitioners. In particular, we integrated feedback from 15 data journalists from daily regional press (Ouest France, Le Parisien, L'Est Républicain...) and from students of the Master's degree in Journalism and Digital Media of the University of Lorraine, where the model was first tested [Chagnoux et al., 2021].

The resulting model is depicted in Figure 4.1, using UML class diagram notation, but omitting class properties for readability purposes.

The organization of the model in 4 layers, reflects the transition from raw facts to the visuals communicated to the audience of the data narrative. On their way to the audience, the facts traverse:

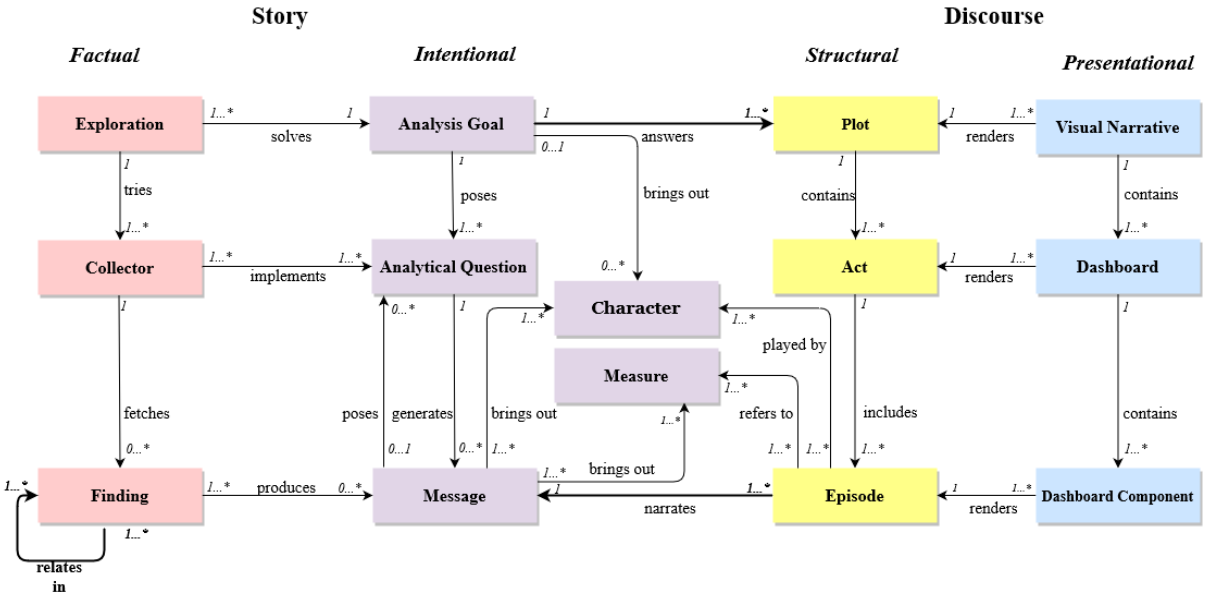


Figure 4.1: The conceptual model for data narratives (relations in bold were extended w.r.t. the original version in [Outa et al., 2020b])

1. **Factual layer**. The factual layer models the *exploration* of facts (i.e., the underlying data), via a set of *collectors* that allow for manipulating facts with varied tools and fetching *findings*[5], in an objective way. *Findings* emerging from explored facts are candidates for participating in the story.
2. **Intentional layer**. The intentional layer models the substance of the story, identifying the *messages*, *characters* and *measures*[6] that the narrator intends to communicate, and tracing how they are obtained through *analytical questions*, according to an *analysis goal*.
3. **Structural layer**. The structural layer models the structure of the data narrative, organizing its *plot* in terms of *acts* and *episodes*. An act corresponds to a major piece of information and a major part of the plot, composed of several episodes. An episode is the granular part of the plot, which conveys a message.
4. **Presentational layer**. The presentational layer models the rendering of the data narrative, i.e., a *visual narrative*, that is communicated to the audience through visual artifacts (*dashboards*[7] and *dashboard components*).

The interested reader is redirected to [Outa et al., 2020b] for a deeper presentation of the model. Here, we will highlight the main decisions behind the model that are necessary for grasping its essence.

Importantly, it should be noted that the concept of *message* is the model's corner stone, which is clearly evidenced by the way we have related message to the other concepts. Essentially, a specific message is rooted in the facts analyzed, conveying essential findings, potentially raising new analytical questions. While a finding can be a pattern like a peculiar value, an association rule, or a path in a decision tree, a message, on the other hand, is the answer to an intentional question that exploits a finding.

[5]Remark that our model of data narrative is agnostic of a specific data model; all the specific details on how facts are collected and support the extraction of findings are encapsulated in the *collector* entity.

[6]Characters and measures are important constituents of a message, indicating relevant elements of the story (as in [Chatman, 1980]), respectively, relevant entities and relevant figures.

[7]We use the term dashboard since it is general enough to accommodate various types of visualizations, e.g. a Business Intelligence dashboard, an infographics, a section in a python notebook, a blog or web page.

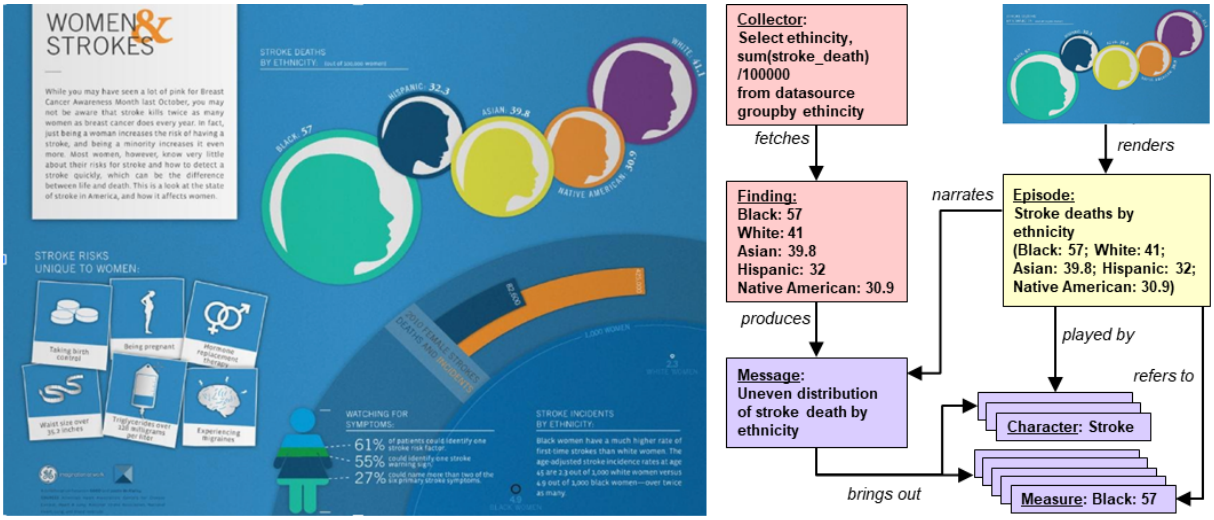


Figure 4.2: Example of data narrative (left) and a partial object diagram for a particular message (right).

Messages allows introducing episodes, the building blocks of the discourse. Each episode is specifically tied to a message which it aims to convey. The relationship between messages and episodes is the basis for structuring stories that address analysis goals, narrated by structured discourses (with cohesive acts being the backbone of the narrative structure) and dashboards their presentational counterpart.

We also point that the factual layer can be thought of as the "objective" one, describing the work around data exploration and model construction, while the intentional layer reflects the "subjective" editorial work of pre-processing findings to turn them into messages.

4.2.3 Example

This subsection illustrates the components of the proposed model using a simple data narrative about women and strokes, described in Appendix B. For illustration purpose, we describe a plausible process for defining analytical questions and collecting data, which is not precised by the narrator.

The final result, a *visual narrative* is depicted in Figure 4.2 (left side), taking the form of an infographic. The *plot* warns women about stroke risks by combining diverse information about risks, symptoms and incidents. The plot is organized in a unique act and six episodes, each episode narrating a message (listed in next paragraph). This act is rendered with a *dashboard* displaying complementary visual information. Six *dashboard components* render the six episodes. For instance, the top right corner of the dashboard displays stroke deaths by ethnicity. Visual artifacts (in this case, circle sizes) are used for carrying the message (here, putting in evidence that black women are the most impacted by stroke deaths).

We summarize the *messages* in the example, from top-left to bottom-right: (m_1) the overall situation of women's stroke in the USA, (m_2) the uneven distribution of stroke death by ethnicity, (m_3) the risks unique to women, (m_4) the rates of women stroke deaths and incidents, (m_5) the poor ability of patients to identify symptoms, and (m_6) the impact of ethnicity in stroke incidents.

Typically, a data narrative starts with an *analysis goal* and a set of *analytical questions*, reflecting the narrator's intention. Here, the narrator's analysis goal is to narrate facts about women and strokes in the USA. An example of analytical question is: Which characteristics of women (age, ethnicity, weight, etc.) have an impact on stroke deaths? Message m_2 answers this question, evidencing that ethnicity is a critical factor. It brings out ethnicity as a *character*, i.e.,

a relevant entity of the story. Analogously, the ratios by ethnicity are brought out as relevant *measures*, i.e., relevant figures in the story. We can note here that characters may appear in several episodes, esp. the main cast (e.g. women, stroke), while others are only supporting in an episode (e.g. symptoms).

A data *exploration* is built by the narrator, who called several *collectors* for analysing data and collecting *findings* in order to answer analytical questions. For example, a collector may query a dataset of female patients in the USA, asking for stroke deaths by ethnicity. The ratios of stroke deaths by ethnicity constitute a finding that supports message m_2, stating the uneven distribution of stroke deaths by ethnicity (black women being the most impacted).

Figure 4.2 (right side) illustrates a partial object diagram concerning message m_2, from the collection of findings to the rendering of an episode.

4.2.4 Experiments and results

In this section we describe our experiments to validate the model and discuss some lessons learned.

Protocol. We conduct two types of experiments aiming to investigate whether: (i) the concepts of our model appear in existing data narratives (and in the description of the crafting processes, when available), and (ii) the main concepts of such data narratives are included in our model.

For the former, we defined a reengineering method, which looks for traces of model concepts (e.g. analytical questions, messages) across the visual rendering of a data narrative or its description. We manually analyze several online data narratives following this method. Details and algorithms are described in [Outa, 2023].

For the latter, we implemented a proof of concept web application[8] helping a narrator in the crafting of a data narrative while interactively exploring a database [Outa et al., 2020a]. We use this application for showcasing that existing data narratives can be crafted based solely on the model concepts.

A large review on the literature about the practical implementation of data narrative concepts (see Appendix A.4 in [Outa, 2023]) completes these experiments.

We remark that the experiments and user studies conducted to validate our proposal of data narration process (to be described in next section), indirectly also validate the subjacent conceptual model.

Lessons learned. As a result of the reengineering experiment, we managed to successfully identify and delineate the key concepts embedded in the data narratives as well as the corresponding elements in their crafting descriptions. This identification of concepts not only confirms the model theoretical foundations, but also highlights its applicability in practical real-world scenarios.

The web application allowed to craft data narratives mimicking the original ones, of course, with very simple visual artifacts. Importantly, all concepts could be recreated, which is an encouraging result, even if our experiments are small enough to banish any attempt of completeness claim.

[8]The code is available on Github: `https://github.com/OLAP3/pocdatastorytelling`

4.2.5 Discussion

This section introduced a conceptual model for data narrative, by extending a classical model of narrative [Chatman, 1980] to reflect the transition from raw data to the visual rendering of messages derived from data analysis. Our model translates fundamental concepts of narration to their respective counterparts when it comes to data narration and involves the collection of data, the extraction of key findings and the corresponding messages to the audience, the structuring of these findings and the ultimate presentation via visual -or other- means via a set of dashboards.

We showcased the model through several real examples and implemented a proof of concept web application helping a narrator devising a data narrative while interactively exploring a database. While for now it can only be used to craft simple narratives, this prototype can be the basis for the creation of more sophisticated ones, once more collectors, dashboard components and dashboards are implemented.

More generally, we found that the conceptual model proved to be an effective tool for communication and co-construction in a transdisciplinary context (the Madona project). In particular, its intuitive form enabled to involve all participants and served to establish consensus during debates. It made it possible to compare the modeling with the ground by verifying that important concepts are not omitted. Furthermore, journalists indicated that *the formalization work, despite its complexity, allowed them to better understand their practices and will prove very useful during the training of future journalists.*

The spirit of the model is to be general enough to accommodate to different application contexts. Indeed, the model can be refined to cope with the particularities and common practices of a given domain, or even to the preferences and intentions of a given narrator. As a proof of concept in this direction, we refined some parts of the model to fit the particularities of the OLAP context [Vassiliadis et al., 2024]. In particular, we take advantage of the multidimensional representation of the underlying data to provide a richer representation of findings (the extended concept is called *highlight*), leveraging dimension members (possible characters of the story) and indicators (possibles measures of the story). Further concepts representing typical OLAP analysis behavior are also modeled. A richer representation of messages is also proposed in [Outa, 2023], tighty relating highligths and messages in OLAP context.

Other researchers also proposed extensions to specific contexts. In particular, [Calegari, 2022] proposes a model-driven approach to generate data narratives rendered in HTML and Jupiper notebooks, and [Wang et al., 2021] adapted the model to a database teaching context to generate narratives as explanations of query execution plans, rendered in natural language. We hope that our model can inspire other usages, even beyond data narration.

In next section, we go a step forward in the modeling of data narratives, by considering their dynamic aspects, and proposing a process model.

4.3 Data narration process

Foreword

This section summarizes part of the PhD thesis of Faten El Outa, co-supervised with Patrick Marcel, and part of the PhD thesis of Raymond Ondzigue Mbenga, co-supervised with Thomas Devogele and Edgar Brice Ngoungou. It also concerns research collaboration with Panos Vassiliadis.
The proposal was published at ADBIS [Outa et al., 2022] and extended at Information Systems Frontiers [Outa et al., 2023]. An instantiation to the Health domain was published at DARLI-AP [Ondzigue Mbenga et al., 2022a].

This section presents a comprehensive and well-founded process that (i) covers the whole cycle of data narration, from the exploration of data to the visual presentation of the narrative, (ii) accommodates a wide range of practices observed on the field, and, (iii) is founded on a conceptual model of the domain that clarifies the concepts involved in the process.

Considered narratives. We experiment on several data narratives described in Appendix B, namely, Climate, Tennis, Covid and Tuberculosis.

Next subsections motivate and describe the proposed process.

4.3.1 Review of literature and practice

We review the literature on data narration processes, and we analyze a survey with data journalists [Chagnoux, 2020] in order to understand how they craft a data narrative.

As an outcome of the former, we find that most of the works describing the data narration process agree on 3 main phases: *analyzing* (to retrieve findings), *structuring* (organizing the information gathered into narrative pieces) and *presenting* (crafting visual artifacts). Automated data narration is still in its infancy, mainly applying rigid patterns and lacking the necessary flexibility of moving between the 3 phases.

One of the key findings is that the intentional layer of the model presented in Figure 4.1 is largely absent from the works reviewed. This means the substance of the story, i.e., the narrator's composition of story elements (analytical questions, messages, etc.) is ignored. We claim that this absence is regrettable; if data narratives are to be shared, reused, and have their crafting process documented, then this intentional layer deserves more attention.

Apart from the bibliographical study, the conducted survey allows us to observe the crafting workflows regularly followed by 18 data journalists, and to contrast them to the literature.

It turns out that journalists follow different paths when crafting a data narrative, with a preponderance of activities pertaining to the factual and intentional layers. They enter the workflow either through factual activities, i.e. by exploring a dataset, or through intentional activities, having at least a vague idea of the subject. After this, the workflow becomes mostly linear, with some movements between factual and intentional activities. Usually, data journalists start writing their articles once the analyzing phase is over, and there is no backtrack once the presenting phase is entered. Notably, the journalists attach less importance to structuring activities. At the exception of one of them, structuring activities are either hidden in writing activities or even not mentioned explicitly. Precisely, many of them agree that while data exploration usually takes long, visual storytelling can be extremely fast, potentially done on the fly, with some of them actually not even involved in the writing of the article.

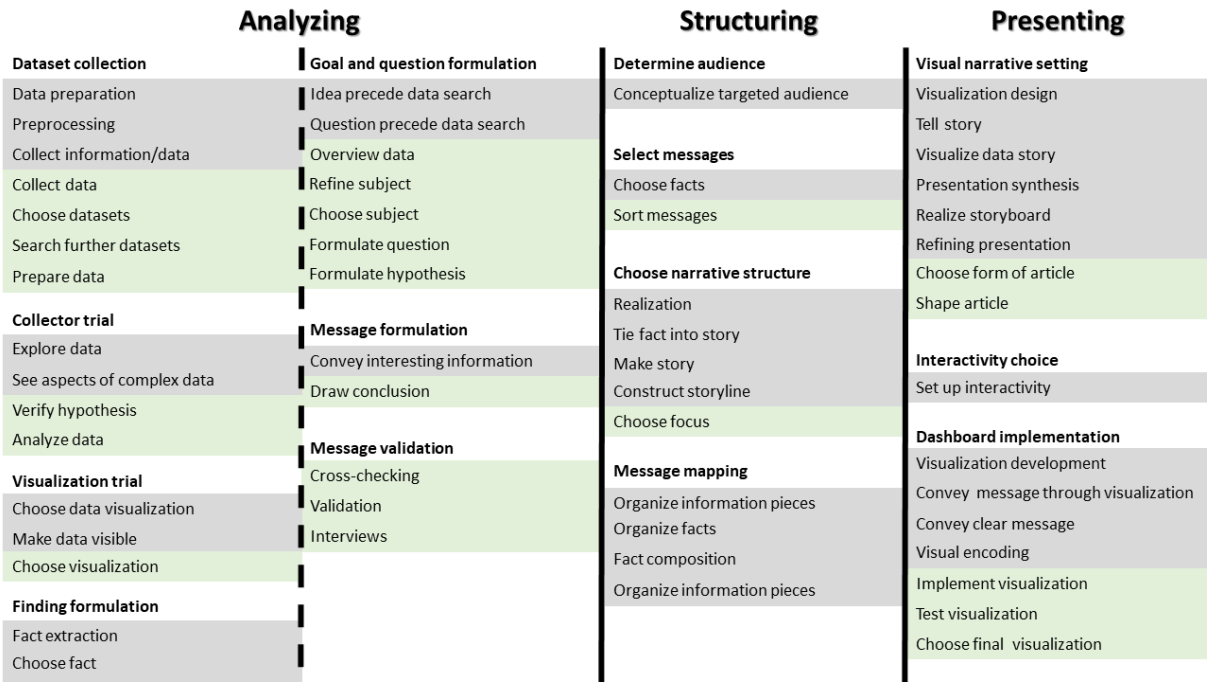


Figure 4.3: The main activities for data narration identified from the literature (in gray boxes) and a survey with data journalists (in green boxes)

The activities described in the literature and by the journalists are listed in Figure 4.3; further description and references can be found in [Outa et al., 2023].

The chasm between literature and practice. Overall, we can say that there is a chasm between what practitioners do and what literature suggests –and in fact, there are deficits in both sides. On the one hand, compared to what is reported in the literature, the work of the data journalists is over-emphasizing the intentional part and under-investing on the structural and presentational parts. On the other hand, when it comes to the literature, the presented methodologies overemphasize presentation and (to some extent) structuring, and pay much less attention to the intentional part. A process that gracefully hosts all aspects of narrative construction would facilitate data narratives that are richer and more intuitive.

4.3.2 Process description

From the literature review and the survey with data journalists, we synthesize a set of requirements for a comprehensive data narration process, and we propose a process that fulfills them. Concretely, a comprehensive process should satisfy the following requirements:

- (R_1) cover the activities and the paths identified by the survey with data journalists, reflecting the intention of the data narrator,
- (R_2) cover the activities of the three phases identified from the literature,
- (R_3) allow the free back-and-forth transition between phases,
- (R_4) clearly delineate the different layers of the conceptual model within its activities.

We propose a comprehensive data narration process that covers the stated requirements.

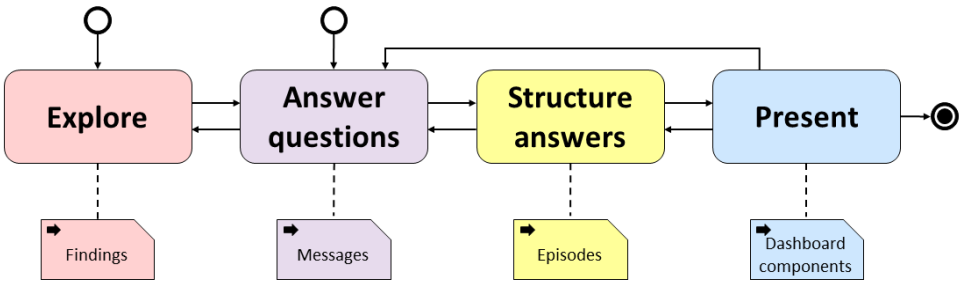


Figure 4.4: The data narration process

The phases of the process are illustrated in Fig. 4.4. All phases are accompanied by the resulting outcomes, which are exactly the basic constituents of our conceptual model (Requirement R_4). Note that, the incomes of the *structure answers* and *present* phases are more than just the basic constituents; rather, they are the organization of episodes and dashboard components. We retain the same coloring (pink for factual exploration, purple for intentional question-answering, yellow for structuring, and blue for presentation).

We remark that the factual and intentional layers of the conceptual model are well differentiated here, contrarily to the literature that mix them into one phase.

Consistently with journalists practices, the process flexibly starts either with the existence of a data set to be explored, or with the emergence of an initiating question to be answered. This flexibility is important in the sense that prescribing a specific starting point for the collection of findings from the data is not what practitioners typically do. The internals allow the flexibility of exploring several paths, that can be chained according to narrator's habits and the specificities of the task on hand, alternating exploration of data, answering questions by deriving messages, structuring the answers and presenting visually the structured answers (Requirement R_3).

In any case, the answering of analytical questions, in terms of messages and their formulation, is a task that is practically absent from the related literature, significantly present in the everyday work of practitioners, and structured in our model for the first time.

The following paragraphs present the activities associated with each phase, which are also sketched in Figure 4.5. These activities are abstracted from the literature and survey results (Requirements R_1 and R_2). Note that such activities should not be considered as steps to be executed sequentially. Conversely, many activities can be initiated and executed in parallel, and many activities are frequently performed asynchronously. The arrows in Fig. 4.5 indicate a *depends on* relationship. For example, message validation depends on message formulation, as it is necessary to formulate messages before validating them. In addition, at any time, it is possible to come back to previously executed activities (e.g. to rewrite messages or formulate new ones). Backtrack arrows are omitted for clarity.

We remark that a new activity, *act and episode writing*, is added to explicitly state the task of conceiving, naming, annotating and contextualizing episodes and acts. In this way, the plot of the data narrative is produced. This activity materializes the concepts of acts and episodes depicted in the conceptual model, which are implicit both in the survey and the literature.

Explore. The explore phase, handling the factual layer, concerns several activities: (i) dataset collection, concerning source selection, data extraction, integration and preprocessing, (ii) trial and reuse of several collectors (i.e. querying, profiling and mining tools) and (iii) trial of diverse visualizations (crosstabs, graphics, clusters, etc.) for collecting findings, then, (iv) finding formulation, concerning the expression of findings and their relationships, and (v) finding validation,

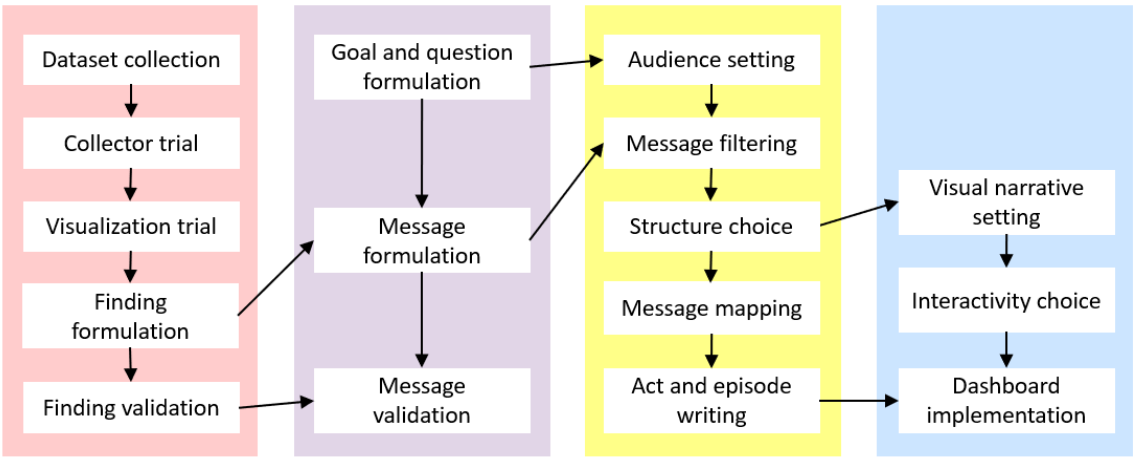


Figure 4.5: Activities for data narration (→ indicates a depends on relationship)

which is typically done via statistical tests and crosschecking. Note that some findings may lead to additional analysis, triggering more collectors and visualisations, or even the collection of more datasets. This phase is time-consuming; data journalists measure it in days or even in months.

Answer questions. This phase handles the intentional layer and concerns activities for (i) formulating goals and questions, (ii) drawing messages from findings, and (iii) validating messages. It supports explicitly the data narrator's intention, as its proposed activities help in formulating an analysis goal and a set of analytical questions that reflect their intention. Furthermore, to cope with literature lacks (evidenced in Fig. 4.3), we propose a message formulation activity, concerning the derivation of messages from findings, and the identification of characters and measures to be highlighted to the audience. We remark that while finding validation is typically done against data (statistical tests, crosschecking), message validation concerns human tasks, as interviews with experts (as done by data journalists [Chagnoux, 2020]) and comparison with the state of the art (as done by data scientists [Ondzigue Mbenga et al., 2022a]).

Structure answers. This phase handles the structural layer, describing activities for organizing the plot of the data narrative in terms of acts and episodes. Plot setting starts by (i) determining the audience, (ii) eventually selecting a subset of messages for such audience, and (iii) choosing an appropriate narrative structure. Then, (iv) messages are mapped to acts and episodes, and in turn, (v) acts and episodes are written. The result of the structuring is an *episode*, which is the annotation of a message with comments on the context, significance, essence, etc., in other words with the content that makes the message interpretable by human beings. Also, observe in Fig. 4.5, the existence of a specific activity to make the actions of writing acts and episodes explicit. The activities of this phase can be performed before or at the same time as choosing visual means.

Present. Finally, the present phase handles the presentational layer, and includes activities for (i) setting the type of visual narratives, (ii) setting the interactivity mode, and (iii) implementing dashboards for conveying acts and episodes to the audience. Such activities carry on the visualization level and build for each act an associated dashboard and present the narration in a complete visual narrative. Remember that *dashboard components* are representations of episodes in (typically) a visual form of communication, including text, figures, charts, data plots, or any other means to convey the message.

In [Outa et al., 2023], we detail the workflow for the answer questions phase, the one being neglected in the literature. It covers the activities and paths reported by data journalists, while also being founded upon and coherent with the conceptual model. Several paths are added based on discussions with data journalists and observations of many data narrators.

4.3.3 Data narration scenarios

The proposed process allows the free back and forth transition between phases (Requirement R_3), some paths being more typical in specific situations. This subsection presents several examples of such situations, representing some common unfolding scenarios described by practitioners or observed. Scenarios are identified based on the following: the study about data journalist practices described in Subsection 4.3.1, the analysis of several data narratives and their associated processes published by data journalists (described in Appendix B.2), and the observation of several practitioners (as will be detailed in Section 4.3.5). These scenarios are sketched in Figure 4.6 by means of regular expressions.

An *exploratory* scenario is commonly observed when the narrator does not have in-depth knowledge of the datasets. It represents situations where the narrator only has a vague idea of the analysis goal (or no goal at all), where many iterations of questions-explorations are necessary to formulate and answer clear questions. This scenario contains many activities and transitions between the phases of *explore* and *answer questions*. Once the exploration is completed and messages are validated, next activities can be linearly performed to structure and then present the data narrative. A good example of this scenario is a data journalist's notebook describing the process followed to build a data narrative about covid pandemic in a French region (described in Appendix B). In this notebook, the data journalist shows the effort put in the many iterations to collect, clean and explore the data and highlights the formulation and validation of messages.

A *pre-canned* scenario corresponds to crafting processes where goals and questions are well defined from the beginning. It is typically observed for periodic or repeated studies, looking for well-known patterns, for example, reporting the results of an election. In this scenario, phases are chained quite straightforwardly, with no need to come back to precise questions or refine collectors. The structure and presentation are typically reused. As an example of this scenario, see a series of data narratives about legislative elections at Rue89Strasbourg Newspaper[9].

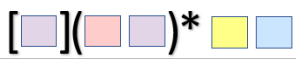
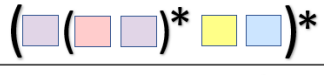
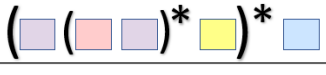
Scenarios	Regular expressions for crafting data narratives
Exploratory	
Pre-canned	
Question-by-question	
Delegated-presentation	

Figure 4.6: Regular expressions representing the unfolding of phases in different data narration scenarios. Colored boxes represent phases, respectively pink for Explore, purple for Answer questions, yellow for Structure answers and blue for Present.

[9] https://www.rue89strasbourg.com/author/raphaeldasilva

A *question-by-question* scenario consists in chaining all phases, one question at a time. In a loop, for each question, an exploration is launched in order to find one or several messages that answer this question. Then, these messages are structured and presented in the rendered data narrative before proceeding with a new question. This scenario concerns more back and forth transitions among all phases. We observed this scenario with beginners, who tried to order and present messages just after their formulation before posing new questions. Students can even go message by message. On the contrary, professionals tend to express most analytical questions at an early stage.

A *delegated-presentation* scenario corresponds to professional environments where the presentation phase is delegated to a specific team at the end of the process. There can be (or not) some iterations among the previous phases, preparing the plot. This scenario was reported by several interviewed data journalists [Chagnoux, 2020].

4.3.4 Instantiation to the Health domain

In previous subsections we proposed a process model and we illustrated that it is general enough to accommodate to several data narration scenarios and practices. In this section we go a step forward, showing how the model can be instantiated to a particular application context.

To this end, we describe the crafting of a data narrative about tuberculosis pandemic in Gabon. This narrative is intended for public health authorities and experts in Epidemic Intelligence (EI), with the goal of describing the epidemiological situation of tuberculosis in a pilot area of study, the Libreville-Owendo-Akanda health region, before a countrywide move.

The crafting process customizes the general process described in this section, by incorporating specific features of epidemiology, best practices in EI, and communication to epidemiologists and public health authorities. In particular, data collection, preprocessing and analysis have a key place in the process, such tasks being at the core of EI. In addition, classical statistical analysis is enriched with other data mining tasks and confronted to the state of the art, the latter being a specific requirement when addressing to a scientific audience. From a technical perspective, the underlying system supports spatio-temporal data of very heterogeneous quality.

Instantiated process. The instantiated process is sketched in Figure 4.7. It reuses and adapts the 4 phases of the general model (although named differently), the initial phase being the intentional one. Indeed, in EI, the analysis goal and many analytical questions must be well defined from the beginning.

We remark, that even if the process allows back transitions, both, phases and activities inside phases, are presented sequentially, which is very intuitive and follows the same organization that other EI processes and protocols, to which EI analysts (the actual narrators) are used to. For this same reason, the *message formulation* activity is included at the end of the *data exploration* phase, to avoid a back transition.

Some activities are split (e.g. *goal and question formulation*) and many others are merged (e.g. *interactivity choice* is included in *visual narrative setting*, and *message mapping* is included in *structure choice*). In this way, *act and episode writing* is merged with *dashboard implementation*, as EI analysts prefer to solve them together. But the main changes happen inside the *data exploration* phase. Indeed, many activities were added to evidence main EI analysis steps (e.g. *result interpretation*) and to introduce EI specific tasks (e.g. *epidemic risk evaluation*).

Highlights. We discuss here the main lessons learned during the instantiation of the process. First, statistical data analysis is not sufficient for public health decision making. A systematic comparison with the state of the art, by comparing the figures obtained, is imperative in order

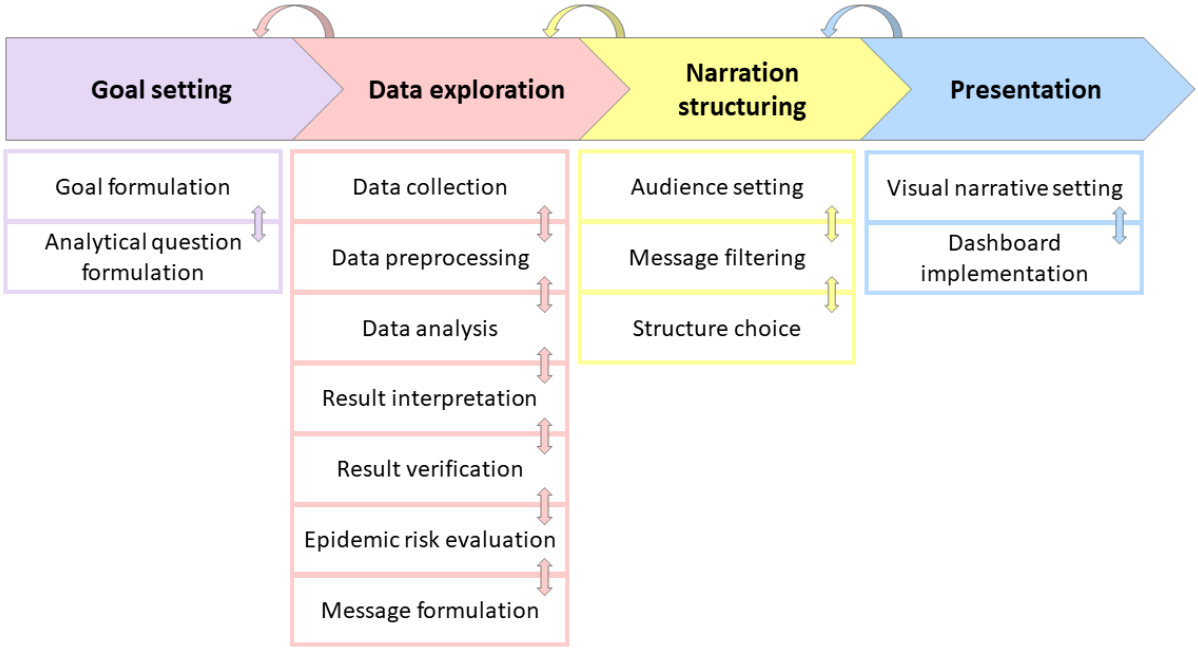


Figure 4.7: Data narration process adapted to Epidemic Intelligence context

to discern global phenomena from regional or seasonal peculiarities. Thus, decision-makers can judge which dimensions of the patient profile are in agreement with the situation in other countries, for which joint actions can be put in place, and which relate to the studied population. Similarly, the results obtained must undergo extensive testing in order to prove its statistical value. As the target audience is predominantly scientific, these results can be communicated.

Second, unlike pre-canned scenarios, in scientific narratives analytical questions are not all known in advance. On the contrary, new questions may arise during data analysis. For example, the distribution of tuberculosis cases by treatment outcome revealed that 74.60% of patients are lost to follow-up. This alarming finding led to a new analytical question asking "what are the epidemiological characteristics of patients lost to follow-up?". Iterations between *goal setting* and *data exploration* phases are often necessary. New findings can also impact previous messages and require updating.

Finally, the data narrative should allow easy navigation between dashboards. In addition, decision-makers like grouped messages about patient characteristics and epidemic indicators, for which a thematic organization (e.g. all messages concerning patients' age in a same act) with navigation links is perfectly suited, and message mapping is done on the fly.

4.3.5 Experiments and results

In this section we describe our experiments to validate the process and report our main findings.

In what follows, we consider several data narratives, either crafted during two challenges that we organized, or crafted by practitioners who also reported the followed processes. All of them are described in Appendix B.

Protocol. Our experiments aim at answering the following questions: (i) Does the process *cover* all necessary activities performed by data narrators? (ii) To what extent do the process *phases contribute* to the quality of the data narratives? (iii) Is the proposed process *consistent* with the reported ones? (iv) Is the instantiated process *adequate* to EI practices?

For the first question, during a challenge in a workshop, we observed several narrators with various profiles while they crafted data narratives for answering the challenge. In particular, we observed whether their actions corresponded to the activities defined in our process.

For the second question, an experienced data journalist assessed the quality of data narratives crafted by Master students during a challenge, and judged the completion of each process phase. Concretely, we investigate the correlation among phase completion and narrative quality.

For the third question, we analyse some published narratives and the associated processes followed by their narrators. Concretely, we investigate whether the proposed process is coherent with the documented ones, highlighting the scenarios that better represent them.

The instantiation to the Health domain is validated by EI analysts and public health authorities during specific workshops and discussed with analysts of neighbour countries.

Coverage. Our main observation is that the proposed process covers the activities (and their chaining) of the 3 teams participating to the challenge, whatever the initial idea, the topic chosen, or the style of visual narrative. In more details, we find that each team struggled at the beginning with the choice of the analysis goal and the datasets to use. In all cases, the first explorations did not return any findings (*finding formulation* arrives a bit later after the trial of several collectors), which did not prevent the teams to continue the crafting. More importantly, the observers note that no activity conducted by the teams is absent from those listed in Figure 4.5.

Interestingly, all teams started with a vague idea of the topic they want to treat, which is refined after many iterations among *dataset collection*, *collector trial* and *question formulation*. This clearly corresponds to an exploratory scenario. Furthermore, we identify some repeated sequences of activities, e.g. *goal and question formulation* followed by *collector trial*, which also illustrate the tight link between *explore* and *answer questions* phases. All teams used a unique timeline for structuring their narratives, which are rendered with varied styles.

We can also note that our proposed process remains tailored for the task at hand. Indeed, the observed activities cover almost all the activities of our process. The remaining ones, pertaining to the *structure answers* and *present* phases, were likely completed after the workshop, as the teams were allowed to continue their crafting during 3 additional days.

Phases contribution to narrative quality. For assessing the relationship between process phases and narrative quality, we asked an experienced data journalist to evaluate data narratives crafted by 44 Master students, assessing both their quality and the perceived phase completion.

Narrative quality is assessed on a scale from 1 (lowest) to 7 (highest), using 3 criteria (previously proposed in [Bar El et al., 2020]): (1) Informativity –How informative the narrative is, and how well does it capture dataset highlights? (2) Comprehensibility –To what degree is the narrative comprehensible and easy to follow? (3) Expertise –What is the level of expertise of the narrator?

The level of completion of each phase (answer questions, structure answers and present), is deduced from the narrative, as the data journalist was not present during the crafting. The data journalist was asked to assess how much of the *answer questions* phase is completed, based on how well the data narratives translate the expression of the intention of the data narrator and how much the subject is investigated. In the same way, the data journalist assessed how much of the *structure answers* and *present* phases are completed. The *explore* phase is omitted from the evaluation because students reported only the rendered data narratives without providing any documentation for the exploration conducted.

	Assessed quality				Perceived completion		
	Info	**Comp**	**Expe**	**Avg$_Q$**	**C$_{ans}$**	**C$_{str}$**	**C$_{pre}$**
Min	1	1	1	1	1	1	1
Max	5	6	5	5.33	6	6	7
Avg	3.38	3.63	3.21	3.43	3.00	3.67	4.17
Stddev	1.13	1.50	1.18	1.14	1.44	1.37	1.52

Table 4.1: Assessed quality (informativity, comprehensibility, expertise, and average quality) and perceived completion (of answer questions, structure answers and present phases) of data narratives of Master students. We report minimum, maximum, average, and standard deviation for each criterion.

The results of the evaluation are reported in Table 4.1, evidencing varied quality and completion. Students were observed during crafting, and some of them, especially those expressing difficulties, are asked to log their sequence of activities. This helps them to start, particularly by writing down the analytical questions that guide the data analysis and the obtained messages.

As to the different phases, the *present* phase is better completed than the two others. In addition, we measure the correlation (using Pearson correlation coefficient) between the average quality (Avg_Q in Table 4.1) and the completion of the three phases. The correlations are, respectively, 0.7 for answer question completion (C_{ans}), 0.85 for structure answers completion (C_{str}), and 0.87 for present completion (C_{pre}). Interestingly, the completion of the three phases is correlated to the overall narrative quality.

We also measured the correlations between the level of expertise and the completion of the three phases, the results being slightly higher for the *answer question* phase (0.79 for answer question completion, 0.77 for structure answers completion, and 0.73 for present completion).

These correlations evidence that the *answer question* phase influence narrative quality at least as much as the other phases, which confirms our claim about its importance for data narration.

Comparison to documented processes. We study four works that documented (at least some portions) of the crafting process followed to produce the Climate, Tennis, Covid and Tuberculosis data narratives, described in Appendix B.2.

For three of them, namely Climate, Tennis and Tuberculosis, the process was clearly described. For analysing them, we just need to match the activities listed by data narrators to those of our process, highlighting the flow of activities. Nevertheless, they described the overall activities accomplished, without detailing every iteration adopted during the crafting process.

For Covid narrative, the process is meticulously reported in a Python notebook. It covers data exploration, with references to goals and questions, but few explicit references to messages. Therefore, we also analysed the visual data narrative for matching messages. The activities of structuring and presenting were not mentioned explicitly by the data journalist.

Figure 4.8 lists the activities performed in the analyzed processes, which are also sketched as a sequence of boxes, colored as the phases of our process.

We find that all the reported processes and activities could be matched to those of our process and the flow between activities is also congruent with our process. In addition, all processes describe many iterations among the initial phases, following an exploratory scenario, even if some of them just illustrate some examples of questions and collectors. Finally, we stress that intentional activities (of the *answer questions* phase) are present in all the reported processes.

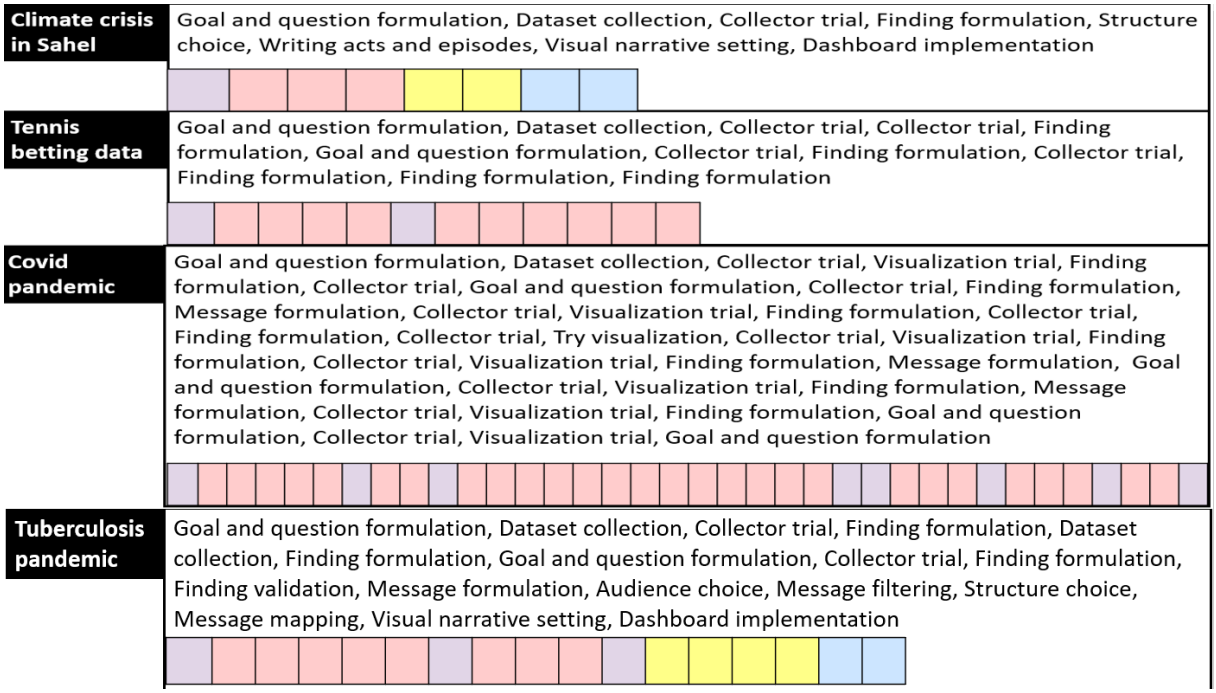


Figure 4.8: The activities of documented processes created by various skilled data narrators.

Instantiation to the Health domain. An interactive data narrative, composed of interconnected interactive dashboards, was presented to EI analysts and public health authorities during specific workshops, detailed in [Ondzigue Mbenga, 2023].

Feedback is very positive, and firstly concerns the messages themselves (some of them alarming about unexpected factors), but also the thematic structuring of the plot, that perfectly matches their needs, and the choice of data visualizations.

They stressed that importance of the geographic dimension for assessing the spatial and spatiotemporal extent of health problems. The restitution in the form of maps is to be favored, but also the spatial correlations. The latter could be incorporated for next workshops.

We also distinguished different profiles among the decision-makers. On the one hand, data analysts need interactive navigation, but on the other hand, public health authorities need a more comprehensive and guided reading of the narrative. The challenge is to find a good balance for rendering, both guided and interactive. Thus, two versions of the data narrative were implemented, with different visual rendering: (i) an interactive narrative, and (ii) a video, capturing a particular navigation through the interactive narrative, with audio explanations.

Finally, the instantiated process was also presented to peer data analysts of neighbour countries [Ondzigue Mbenga et al., 2021], resulting in nice feedback and reuse opportunities.

4.3.6 Discussion

This section proposed a data narration process that covers the whole cycle of data narration, from data exploration to the visual presentation of the data narrative. Importantly, the process reflects the intention of the data narrator by incorporating activities covering the formulation of their goals, questions, and messages.

Backed by a literature review and a survey with data journalists, the process accommodates a wide range of practices observed on the field, via clearly delineated activities, while being well founded upon a conceptual model of the domain. Indeed, the comparison of the different versions

of the process with the field (observation and feedback) made it possible to propose more complex and complete workflows, either by integrating new activities (for example, the frequent recourse of journalists to experts to validate stages of their work), or by trying to reconcile the linearity of the model with the proliferation of data investigation work. Furthermore, several user studies and the comparison to reported practices allowed to validate both, the process consistency with respect to common practices and the completeness of its activities.

Even if the process is general enough to accommodate varied practices, we can expect that some refinements could be necessary to adapt to specific application contexts. We experienced this in the Health domain, developing an instantiated model that maintains most of the activities of the general process, adds some activities specific to the domain (e.g. epidemic risk evaluation) but above all, simplifies many paths in the workflow, to be less general and more easily understood by practitioners. We hope the process could be also reused in other contexts.

4.4 Conclusion

This chapter presented our contributions for understanding and modeling data narratives.

We first studied static aspects of data narration, and proposed a conceptual model of data narratives, translating fundamental concepts of narration theory to the fields of data exploration and visualization. The model is supported both, by a large review of the literature, inventorying related concepts from several domains, and by multiple exchanges with practitioners from several areas, in particular journalism, communication science and business intelligence.

We then incorporated dynamic aspects of data narration and proposed a process model, that covers the whole cycle of data narration. The process model is also backed by a literature review and a survey with data journalists, and accommodates to a wide range of practices observed on the field.

We illustrated a particular instantiation of the conceptual model to the Business Intelligence domain [Outa, 2023], and other researches also extended the model to e-Learning [Wang et al., 2021] and Model-driven Engineering [Calegari, 2022] domains. Likewise, we illustrated the instantiation of the process model to the Health domain. All these cases, practically show the potential of the models to accommodate to different applications and practices.

We believe that both models, static and dynamic, can serve as a stepping stone for future research in the area of data narration, specially for the implementation of tools for guiding the narrator all along the process as well as automating tedious or complex tasks. We indeed believe that holistic approaches to data narration (from exploration to visual presentation) should be adopted, and we particularly insist on the importance of the intentional phase of the process. Activities in this phase (e.g., message formulation, message validation) are likely to be the most difficult to automate. This a clear first step to the development of approaches for data narrative management and sharing.

Our definition of data narrative and our models has been studied in several domains, including: Cultural Heritage [Kadastik and Bruni, 2023], Social Sciences [Risam, 2023], Public Policies [Parker et al., 2023], Business Intelligence [Gunklach et al., 2023a], and Master Data Management [Kuznetsov et al., 2022].

They have also been considered for specific tasks, including: finding of narrative evidences [Nagel et al., 2023], narrative summary [Ghodratnama et al., 2021], plot generation [Ranade and Joshi, 2023], and narrative visualization [Edmond and Bednarz, 2021].

Finally, several surveys on data narratives describe and compare our work [Ranade et al., 2022], [Lezcano Airaldi et al., 2022], [Schröder et al., 2023].

Chapter 5

Conclusions

This chapter concludes the dissertation by summarizing the contributions and discussing future research directions.

It therefore positions ongoing PhD theses, related master projects and ongoing research projects, summarized below. Some perspectives partially rely on reflections from the preparation of a keynote[1] and some tutorials[2][3][4], and the discussions that followed.

Advising, projects and collaborations

PhD theses:
Flavia Serra (started in 2020), co-supervised with Patrick Marcel and Adriana Marotta.
Raphaël Bres (started in 2021), co-supervised with Ana Maria Olteanu Raimond, Cyril de Runz and Arnauld Le-Guilcher.
Guillaume Tejedor (started in 2023), co-supervised with Hélène Blasco, Patrick Marcel and Nicolas Labroche.
Hiba Merakchi (started in 2023), co-supervised with Thomas Devogele.

Postdoctoral project: Louise Parking (2023-2024), co-supervised with Béatrice Markhoff.

Master theses and projects: Quentin Barreau (2021), Valentin Fradet (2022), Jimmy Rata Gobal (2023), Imen Haddar (2023), Boubacar A. Bah (2024), Jules Harrouet (2024).

Research projects:
OPTIMEDIAS - *Optimization of Data Exploitation by Artificial Intelligence in Health*[5] (2022-2025), regional funding.
JUNON – *Digital twins for natural resources*[6] (2022-2027), regional funding.
Data quality within data preparation for Big Data analysis[7] (2023-2026), Uruguayan funding (CSIC).
IntForOut – Multisource spatial data INTegration FOR the monitoring of ecosystems under the pressure of OUTdoor recreation (2024-2027), national funding (ANR).

[1] *From source data to data narratives: accompanying users in the way to interactive data analysis*, keynote at the 2020 ADBIS, TPDL & EDA joint conferences [Peralta, 2020] – Video:
`https://eric.univ-lyon2.fr/adbis-tpdl-eda-2020/adbis-tpdl-eda-2020/author-material/keynote-peralta.mp4`

[2] *Exploratory data analysis: from insights to storytelling*, tutorial at the 8th EGC Winter School on "Humans in the data exploration and learning loop" (é-EGC 2022)

5.1 Synthesis of contributions

This dissertation addressed the general problem of supporting data exploration and narration, by leveraging users' behavior, interests and intentions.

We addressed three specific challenges:

1. qualify data explorations and learn users' analysis behavior from users' past explorations,
2. model and learn users' interests, and
3. model the static and dynamic aspects of data narration.

Chapter 2 addressed the first challenge, presenting our contributions for understanding, modeling and learning the way users analyse data. We first proposed a model for queries and explorations from the prism of users' skills, based on a large set of features describing varied aspects of a query, and its context within the exploration.

We then proposed an approach for qualifying OLAP queries and explorations, modeled respectively as classification and skill acquisition problems, which succeeded to capture the characteristics of the workload, experts' advice and users' skills.

The extension from OLAP to a more complex SQL environment, and avoiding the need of experts for labeling, introduced the challenge of workload segmentation. We proposed three methods for tackling it, using respectively unsupervised (similarity-based), supervised (transfer) and semi-supervised (weak-labelling) learning methods. The proposed methods got good quality and outperform state-of-the-art timestamp-based strategy.

Finally, we proposed an unsupervised approach for learning users' behavior. The proposal is based on a similarity measure tailored for explorations, which paired with a clustering algorithm, allowed the identification of several types of analysis patterns. Our method outperformed state-of-the-art similarity measures for workloads with ground truth, and succeeded distinguishing users' skills.

To our knowledge, our contributions are pioneer on their kind. We showed that learning users' behavior is feasible and is consistent with experts' judgement (when a ground truth is available). We believe that the identification of analysis behavior could be exploited for conceiving more intelligent EDA support tools, for example, for classifying users, personalizing and recommending queries, but also for better understanding users' intentions. The exploitation of analysis behavior in such way is still to be tested, and is discussed in next section.

Chapter 3 addressed the second challenge, presenting our contributions for understanding, modeling and learning users' interests.

[3] *Data Exploration from Insights to Storytelling*, tutorial at the 10th European Big Data Management & Analytics Summer School (eBISS'2022) – Slides:
`https://cs.ulb.ac.be/conferences/ebiss2022/slides/marcel_peralta_1.pdf`
`https://cs.ulb.ac.be/conferences/ebiss2022/slides/marcel_peralta_2.pdf`

[4] *Data Narration for the People: Challenges and Opportunities*, tutorial at EDBT'2023 [Marcel et al., 2023a] – Video (starting at 17:00):
`https://db.disi.unitn.eu/pages/EDBTpedia/sessions/data_narration_for_the_people_challenges_and_opportunitie/`

[5] Original name (in French): *OPTIMEDIAS – OPTIMisation de l'Exploitation des Données par l'Intelligence Artificielle en Santé*

[6] Original name (in French): *JUNON – Des jumeaux numériques au service des ressources naturelles* - `https://www.brgm.fr/fr/programme/junon-jumeaux-numeriques-au-service-ressources-naturelles`

[7] Original name (in Spanish): *Calidad de Datos en la Preparación para el Análisis de Big Data*

We first proposed a two-level framework for developing interestingness measures, consisting respectively of high-level interestingness aspects, and data-oriented assessment algorithms. We showed that even simple measures can help the analysis of users' interests, finding that query interestingness is correlated with exploration quality, that users have not strong preferences for a particular aspect, and interest perceptions change over time.

We then focused on a particular interestingness aspect, relevance, and proposed an approach for learning users' interests in a query workload and recommending relevant queries. We used classification, clustering and recommendation techniques, which succeeded to capture users' interests, being effective in practice, and specially beneficial to novice analysts.

Our recommendation method outperformed state-of-the-art recommenders, showing that a better understanding of users' interests could be beneficial to EDA support tools. Furthermore, the proposal was refined in industrial context and patented [Drushku et al., 2021].

We have not yet investigated the relationship between users' behavior and interests, but among the discovered analysis patterns, some of them clearly indicate users' intentions (specially those patterns showing focused behavior), while other ones translate struggle (e.g. dataset reloader). Also, both users' behavior and interests relate to users' skills, giving opportunities to make novices benefit from experts. We discuss this matters in next section.

Chapter 4 addressed the last challenge, proposing our contributions for understanding and modeling data narratives, where EDA is an important phase.

We first proposed a conceptual model of data narratives, translating fundamental concepts of narration theory to the fields of data exploration and visualization. We then incorporated dynamic aspects of data narration and proposed a process model, that covers the entire cycle of data narration, while highlighting the importance of the intentional phase.

Both models are backed by a large review of the literature covering several scientific domains, and by multiple exchanges with practitioners from several areas, in particular journalism, communication science and business intelligence. We illustrated several instantiations of the proposed models to particular domains, practically showing the potential of the models to accommodate to different applications and practices.

We believe that both models can serve as a stepping stone for the development of tools for guiding the narrator all along the process as well as automating tedious or complex tasks. We indeed believe that holistic approaches to data narration are necessary, and in particular, EDA tools should be better integrated. This vision is also discussed in next section.

5.2 Perspectives

Our perspectives are organized in four groups according to the phases of the data narration process they pertain, namely further support for data exploration (naturally concerning the *explore* phase), enhancement of intentions (centered in the *answer questions* phase and influencing the other ones), switch to audience-driven data narratives (dealing with *structure answers* and *present* phases) and development of an overall data narrative management system. Figure 5.1 positions perspectives with respect to the data narration process described in previous chapter (with some envisioned extensions).

The groups of perspectives are presented starting from the short term ones, including some leads for two starting PhD theses, to the longer term ones, envisioning data narrative management systems and putting the audience in the loop.

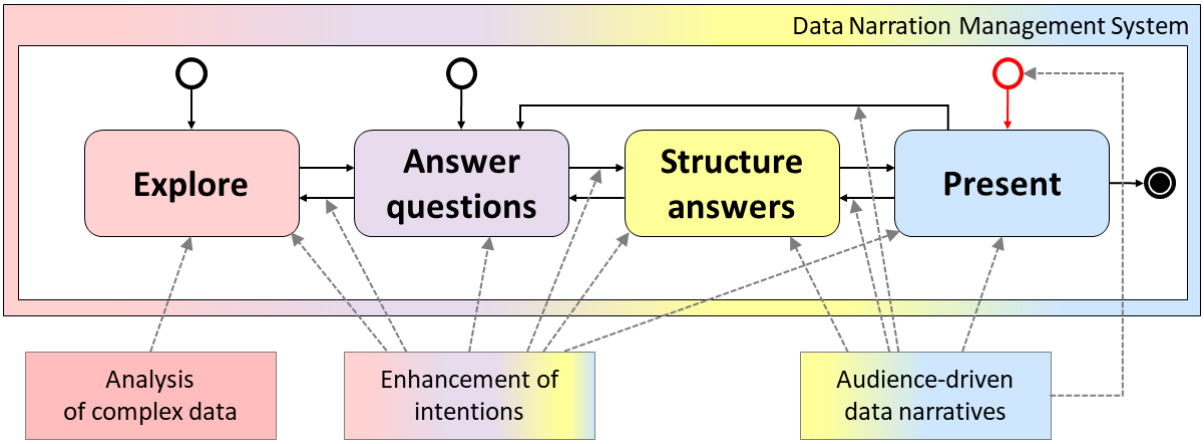


Figure 5.1: Perspectives position w.r.t. data narration process

5.2.1 Analysis of complex data

In Chapters 2 and 3, we presented various techniques for analysis of sequences of complex data (resp. sequences of queries and observations), which allowed to learn analysts' interests, behavior and skills. More generally, applied to sequences of other types of human activities (mobility, lifelong paths, patient records), these techniques allowed the discovery of many patterns of human behavior [Moreau, 2021]. The analysis of sequences revealed itself to be challenging, opening the way for further investigation.

Broaden analysis of sequence similarity. Sequence similarity is a key point for sequence analysis. We proposed two measures, CED (based on Edit Distance, described in Section 2.5.1) and FTH (based on Hamming Distance, described in [Moreau et al., 2021b]). An in-depth comparative study is still to be carried out. In particular, we would like to investigate which characteristics should be favored when choosing and setting the measures parameters, and how to adapt to data quality issues, e.g. to deal with incomplete sequences. These measures should also be reused within a query language allowing the retrieval of similar sequences.

An interesting direction for future work, is a deeper study of the time dimension. For example, we could describe activities according to their start and end timestamps, allowing the detection of richer patterns (e.g. w.r.t. duration and seasonality), and the opportunity for anomaly detection [Boniol et al., 2023]. Activities referring to time points (e.g. a blood analysis) are even more challenging than those pertaining to time periods (e.g. driving). We investigated strategies for time alignment and interpolation [Haddar, 2023], the topic deserving deeper analysis. Of course, the spatial dimension, largely studied both at urban and indoor scales ([Miller, 2017, Kontarinis et al., 2021]), should also being included. We studied the potential of exploiting multiple sensors (e.g. GPS, accelerometer, rotating beacon, break switch) for enhacing sequence semantics [Bisone, 2021]. Putting together activity semantics, time and space is a nice challenge. The PhD thesis of Hiba Merakchi will explore some of these leads.

Exploitation of sequences. An analysis tool, SIMBA, implements our proposal for sequence analysis [Moreau et al., 2020a]. It offers a simple and complete pipeline from raw data to clustering analysis for studying semantic sequences and extracting behavior. Despite the tool proposes a large set of complementary visual indicators (described in Subsection 2.5.2), new indicators are needed for better integrating users' preferences and promoting the explainability

of the process. We approached the problem by studying powerful indicators as fuzzy prototypes [Lesot and Kruse, 2006, Barrau, 2021], but simple indicators for novice users are still more challenging [Soni et al., 2022].

Beyond clustering analysis, the comparison of sequences and the learned knowledge can be exploited for diverse tasks. In particular, the study of sequences of patient records could be used for patient stratification and survival time prediction, topics to be studied in the PhD thesis of Guillaume Tejedor. Patients records are challenging by itself, because of the variety of features (and thus the high dimensionality) typically referring to time points (see discussion about the time dimension above) and their sensibility to quality problems. Our first lead is to look for typical subsequences. In particular, Shapelets [Zhang and Sun, 2023] are good candidates, providing reliability and explainability of the process even to non-experts [Zuo et al., 2019]. The relation of shapelets with survival changing (as studied in [Oubelmouh et al., 2023] for pattern mining), is another interesting lead.

Coming back to EDA support, many use cases can be envisioned. In Chapter 3, we showed how the discovered users' interests can be integrated in a recommender system; an industrial tool exploits the proposal [Drushku et al., 2021]. Similar usages of analysis behavior and skills are still to test. The development of an EDA support tool supporting them is a pending project, ideally to be undertook within an industrial partnership.

5.2.2 Enhancement of intentions

The *answer questions* phase of the data narration process is the one needing more attention. Narrator's intentions spread beyond the expression of an analysis goal and a set of analytical questions. Intentions should guide many (probably all) activities of the data narration process. We sketched this in Figure 5.1 with arrows showing the influence to all the process phases and transitions. We envision many research directions.

A broader model of intentions. We should start by a broader model of intentions. A first attempt was proposed in Chapter 5 of [Outa, 2023], by providing a more detailed model of *message* with detailed connections to findings and analytical questions, and introducing new concepts, as hypothesis and beliefs. But such model needs refinements, which should come from interdisciplinary collaborations, including cognitive sciences and a large palette of practitioners, as done in the Madona project.

From intentions to exploration. We claim that intentions should guide data exploration, even when the analysis goal is unclear. In this sense, [Vassiliadis and Marcel, 2018] envisioned a new EDA paradigm (IAM) supporting intentional querying. They proposed five intentional operators that can be automatically translated to database queries; two of them were implemented [Francia et al., 2022c, Chanson et al., 2022a]. This paradigm deserves more attention, not limiting to intentional operators but conceiving a whole framework around intentions. This includes the recommendation and personalization of intentional queries, but also the reasoning with analysis goals and analytical questions, conceiving messages as answers to those questions, as done by the Goal-Question-Metric approach, proved effective for data quality management [Akoka et al., 2007].

Interesting measures should also be exploited for guiding data exploration, as done in Section 3.3 for relevance. The combination with analysis patterns (as those learned in Chapter 2) is also interesting, as they inform about users' practices but also translate intentions. In this way, we

could discover analysis patterns, relate them to analytical questions, in order to recommend relevant new queries, but also new analytical questions. In this sense, the recommendation of hypothesis for EDA is attracting attention [Nejar de Almeida et al., 2023].

From intentions to structuring and presentation. Narrator's intentions should also guide the discourse. Indeed the narrative structure should consider all the intentional content inside messages, instead of just structuring findings as done in the state of the art. Beside the typical linear and parallel structures, developing intention-based structures in a nice challenge. Furthermore, intentions should also guide the visual choices; dashboard patterns, as those proposed in [Bach et al., 2023], could be extended to follow intentions.

More generally, further effort is necessary for the development of automatic exploration and structuring approaches, producing more complete and complex narratives, but also integrating the narrator's intention in the loop.

A pride of place for data quality. Despite the capital importance of data quality, claimed by practitioners (who report spending months in data collection and curation [Chagnoux et al., 2021]) and the amount of research solutions (data cleaning and quality is within the most frequent research topics addressed in Data Management within the last ten years [Darmont et al., 2022]), data quality management is almost absent in data narration approaches. All automatic tools (e.g. [Wang et al., 2020, Shi et al., 2021]) take as input a unique dataset, supposed to be clean; data collection, integration, transformation and wrangling are left to narrators. Even EDA platforms generally base on the existence of a curated database, typically a data warehouse (e.g. [Youngmann et al., 2022, Muhammad and Darmont, 2023, Lipman et al., 2023]).

We claim that data quality should be a first class citizen in EDA solutions. The analysis of data freshness of cycling routes [Bres et al., 2022, Bres et al., 2023] conducted during the PhD thesis of Raphaël Bres, is an example of how data quality issues impact data management (in this case, cycling network modeling and route recommendation). In this line, within IntForOut and Junon projects, we envision data curation and integration methods, supported by knowledge graphs, and guided by data exploration needs. In the former, EDA will be used for monitoring human activities exerting pressure on biodiversity and ecosystems, in the latter, automatic analysis (with digital twins on water, soil and air usage) will reproduce and predict natural processes.

Going further, data quality should drive not only EDA but also users' intentions handling. A large systematic literature review [Serra et al., 2022b] revealed the importance of contextual aspects for data quality management. In particular business needs and the task at hand frequently condition the quality improvement actions to be applied. The context model [Serra et al., 2022a] and context-aware data quality management methodology [Serra et al., 2023] proposed in the PhD thesis of Flavia Serra, should be integrated to the data narration process. Beyond recommending next queries, recommending curation actions is a challenging topic.

5.2.3 Towards a data narration management system

Our long-term perspective is to develop a data narrative management system, supporting all narration activities in an unified platform. This means considering data narratives as first class citizens, enabling their storage, sharing, reuse and manipulation. In addition to many technical challenges ensuing from the integration of multiple tools (so asked by practitioners), the development of such systems raises many research challenges.

Supporting sharing and reuse. Could we think of data narrative manipulation languages? In a vision paper, we lay the preliminary foundations for data narrative management systems, and introduce a simple logical framework supported by a data narrative manipulation language inspired by the extended relational algebra [Marcel et al., 2023b]. This type of languages, considering all concepts of data narratives, deserve more investigation.

More generally, data narration processes are workflows that should be crafted to be reused. Works on design and management of scientific workflows (e.g. [Cohen-Boulakia, 2022]) could be adapted to the management of data narratives.

A tighter integration of data narration and data exploration. The state of the art shows that data exploration is too much dissociated from narration. We argue to the need to develop an Explore-Narrate-Explore paradigm. The overall idea is to provide partial guidance instead of full-guidance, and let the audience intervene after the presentation of the story. The data narrative is then refined after at each iteration, which reduces the overall time-to-message. This requires to revisit the *answer questions* phase in order to specify Explore-Narrate-Explore intentions, and opens the door for revisiting recommendation approaches to recommend Explore-Narrate-Explore steps.

A tighter integration of data narration and evaluation. Benchmarking data narratives and the underlying crafting process offers many practical usages, in particular for developers of management systems. But what is the quality of a data narrative? In Chapter 2 we proposed methods for assessing the quality of queries and explorations. Shifting to quality assessment of data narratives is still more challenging.

The Data visualization community studied many human- and data-oriented measures, as accuracy, effectiveness, interpretability and user engagement [Lam et al., 2012, Wang et al., 2019, Boukhelifa et al., 2020], and the Data Management community proposed many others, as informativity, comprehensibility, expertise [Bar El et al., 2020] and interestingness (see Section 3.1). Many system-oriented measures come from Data Management benchmarks, as latency or memory-usage. We surveyed and classified quality measures in [Marcel et al., 2023a]. Open questions concern the combination of human-, system- and data-oriented measures, and the proposal of new human-oriented measures (e.g. for textual narration: completeness, readability and conciseness). As the *answer questions* phase has been overlooked, new measures are needed for qualifying intentional aspects, e.g. the completion of the data narrative w.r.t to its purpose.

5.2.4 Audience-driven data narratives

Our last and longer-term perspective is about a better inclusion of the audience in the data narration process. This starts by being able to explain the narration and to conciliate narrator's and auditor's expectations.

Data narratives as explanations. Until now, the purpose of a data narrative is to convey messages describing findings. But what about shifting this purpose to describe *why* and *how* these findings were obtained. Such a shift relies on the ability to explain steps. For example: (*Explore*) How is a finding found? (*Answer Questions*) Why is a particular question posed? (*Structure Answers*) How is a particular ordering chosen? (*Present*) Why is a visual mapping chosen? In this way, the data narrative itself becomes an explanation.

Shift the attention to the audience. To better account for the audience, the data narrative should be adapted to the auditors' profiles (decision markers, data enthusiasts, virtually anyone). Particularly, the discourse (structure, presentation) should change. For example, the choice of structures (e.g. linear vs. non-linear) should enable alternative discourses, the ordering strategy should reflect diversity among the auditors, and message-to-visual mappings should be personalized.

As a first approach, we investigated the impact of context, structure and visual means in message reception and interpretation, through user studies during master projects and internships. We found that users managing a large company (we experienced in a bank) prefer hierarchical structures and normalized visualizations [Fradet, 2022], while within a more varied audience (we experimented with auditors of varied ages and skills), auditors' age and story topic were correlated to auditors' perception of story quality [Rata Gobal, 2022]. We are currently studying the impact of auditors' training and position on dashboard characteristics (structure, color, etc.). The personalization and recommendation of structuring and presentation patterns according to auditors' characteristics is a nice challenge.

Going further, we envision an auditor even starting the data narration process, asking for a data story, sometimes having a well-defined idea of the expected visual narrative. This means being able to start the process by the *present* phase (the red entry point in Figure 5.1) and do many Explore-Narrate-Explore moves, the data narrative being refined at each iteration. Reinforcement Learning has been used to automatically produce explorations (see e.g. [Bar El et al., 2020, Personnaz et al., 2021]), it could be used to automatically produce Explore-Narrate-Explore data narratives.

Collaborative data narration. What we have so far mostly concerns one narrator and one auditor. Eventually we should think of data narrations being crafted by multiple narrators for multiple auditors. Several challenges arise, as reconciling findings and messages, identifying complementary and contradicting viewpoints, and balancing multiple levels of interaction and complexity in visualizations.

We also highlight the need for diversity, inclusion and ethics considerations all along the data narration process, reconciling intersectional[8] points of view. Although diversity and inclusion actions are increasingly considered by Data Management and Machine Learning communities [Amer-Yahia et al., 2023, Amer-Yahia et al., 2022] and ethics considerations are required [Risam, 2023], there are still no initiatives concerning data narration. This is more than just considering multiple auditors, it is a matter of including everyone in the audience.

This dissertation presented several contributions in the domains of Exploratory Data Analysis and Data Narration, undertook in collaboration with several PhD students and colleagues of the University of Tours, as well as associates from other universities, public services and local companies. The presented work opens up numerous and exciting perspectives; some research work (starting theses and projects) have already been launched to address some of these perspectives.

[8]See a nice keynote on *An Intersectional Approach to Data Governance*, by Marie Plamondon at EDBT/ICDT'2023.

Bibliography

[Abadi et al., 2022] Abadi, D., Ailamaki, A., Andersen, D. G., Bailis, P., Balazinska, M., Bernstein, P. A., Boncz, P. A., Chaudhuri, S., Cheung, A., Doan, A., Dong, L., Franklin, M. J., Freire, J., Halevy, A. Y., Hellerstein, J. M., Idreos, S., Kossmann, D., Kraska, T., Krishnamurthy, S., Markl, V., Melnik, S., Milo, T., Mohan, C., Neumann, T., Ooi, B. C., Ozcan, F., Patel, J. M., Pavlo, A., Popa, R. A., Ramakrishnan, R., Ré, C., Stonebraker, M., and Suciu, D. (2022). The seattle report on database research. Commun. ACM, 65(8):72–79.

[Abascal-Mena et al., 2015] Abascal-Mena, R., Lema, R., and Sèdes, F. (2015). Detecting sociosemantic communities by applying social network analysis in tweets. Soc. Netw. Anal. Min., 5(1):38:1–38:17.

[Abelló and Romero, 2018] Abelló, A. and Romero, O. (2018). Online analytical processing. In Liu, L. and Özsu, M. T., editors, Encyclopedia of Database Systems. Springer, 2nd edition.

[Abiteboul et al., 2018] Abiteboul, S., Arenas, M., Barceló, P., Bienvenu, M., Calvanese, D., David, C., Hull, R., Hüllermeier, E., Kimelfeld, B., Libkin, L., Martens, W., Milo, T., Murlak, F., Neven, F., Ortiz, M., Schwentick, T., Stoyanovich, J., Su, J., Suciu, D., Vianu, V., and Yi, K. (2018). Research directions for principles of data management (dagstuhl perspectives workshop 16151). Dagstuhl Manifestos, 7(1):1–29.

[Abiteboul et al., 1995] Abiteboul, S., Hull, R., and Vianu, V. (1995). Foundations of Databases. Addison-Wesley.

[Acar and Motro, 2004] Acar, A. C. and Motro, A. (2004). Why is this user asking so many questions? explaining sequences of queries. In DBSec'2004, Sitges, Spain.

[Akleman et al., 2015] Akleman, E., Franchi, S., Kaleci, D., Mandell, L., Yamauchi, T., and Akleman, D. (2015). A theoretical framework to represent narrative structures for visual storytelling. In Bridges'2015, Maryland, USA.

[Akoka et al., 2007] Akoka, J., Berti-Équille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué-Thion, V., Kedad, Z., Nugier, S., Peralta, V., and Cherfi, S. S. (2007). A framework for quality evaluation in data integration systems. In ICEIS'2007, Funchal, Portugal.

[Ali, 2022] Ali, H. (2022). Pattern-driven analysis of pedestrian movement. Master's thesis, Technische Universität Wien, Austria.

[Aligon et al., 2014a] Aligon, J., Boulil, K., Marcel, P., and Peralta, V. (2014a). A holistic approach to olap sessions composition: The falseto experience. In DOLAP'2014 (CIKM workshop), Shanghai, China.

[Aligon et al., 2015] Aligon, J., Gallinucci, E., Golfarelli, M., Marcel, P., and Rizzi, S. (2015). A collaborative filtering approach for recommending OLAP sessions. Decis. Support Syst., 69:20–30.

[Aligon et al., 2011] Aligon, J., Golfarelli, M., Marcel, P., Rizzi, S., and Turricchia, E. (2011). Mining preferences from OLAP query logs for proactive personalization. In ADBIS'2011, Vienna, Austria.

[Aligon et al., 2014b] Aligon, J., Golfarelli, M., Marcel, P., Rizzi, S., and Turricchia, E. (2014b). Similarity measures for olap sessions. Knowl. Inf. Syst., 39(2):463–489.

[Álvarez-Ayllón et al., 2019] Álvarez-Ayllón, A., Palomo-Duarte, M., and Dodero, J. M. (2019). Interactive data exploration of distributed raw files: A systematic mapping study. IEEE Access, 7:10691–10717.

[Amer-Yahia et al., 2023] Amer-Yahia, S., Agrawal, D., Amsterdamer, Y., Bhowmick, S. S., Bonifati, A., Borovica-Gajic, R., Camacho-Rodríguez, J., Catania, B., Chrysanthis, P. K., Curino, C., Darmont, J., Dobbie, G., Abbadi, A. E., Floratou, A., Freire, J., Jindal, A., Kalogeraki, V., Maiyya, S., Meliou, A., Mohanty, M., Omidvar-Tehrani, B., Özcan, F., Peterfreund, L., Rahayu, W., Sadiq, S., Sellami, S., Sirin, U., Tan, W., Thuraisingham, B., Tian, Y., Tözün, P., Vargas-Solar, G., Yadwadkar, N. J., Zakhary, V., and Zhang, M. (2023). Diversity, equity and inclusion activities in database conferences: A 2022 report. SIGMOD Rec., 52(2):38–42.

[Amer-Yahia et al., 2022] Amer-Yahia, S., Bonifati, A., Favre, C., Fromont, É., Labroche, N., Melançon, G., Sèdes, F., Soulet, A., and Termier, A. (2022). Diversity and inclusion activities in EGC - A 2022 report. SIGKDD Explor., 24(1):52–56.

[Aufaure et al., 2013a] Aufaure, M., Cuzzocrea, A., Favre, C., Marcel, P., and Missaoui, R. (2013a). An envisioned approach for modeling and supporting user-centric query activities on data warehouses. Int. J. Data Warehous. Min., 9(2):89–109.

[Aufaure et al., 2013b] Aufaure, M.-A., Kuchmann-Beauger, N., Marcel, P., Rizzi, S., and Vanrompay, Y. (2013b). Predicting your next olap query based on recent analytical sessions. In DaWaK'2013, Prague, Czech Republic.

[Ba et al., 2014] Ba, C., da Costa, U. S., Ferrari, M. H., Ferré, R., Musicante, M. A., Peralta, V., and Robert, S. (2014). Preference-driven refinement of service compositions. In CLOSER'2014, Barcelona, Spain.

[Bach et al., 2023] Bach, B., Freeman, E., Abdul-Rahman, A., Turkay, C., Khan, S., Fan, Y., and Chen, M. (2023). Dashboard design patterns. IEEE Trans. Vis. Comput. Graph., 29(1):342–352.

[Baeza-Yates and Ribeiro-Neto, 2011] Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). Modern Information Retrieval - the concepts and technology behind search. Pearson Education Ltd., 2nd edition.

[Bar El et al., 2020] Bar El, O., Milo, T., and Somech, A. (2020). Automatically generating data exploration sessions using deep reinforcement learning. In SIGMOD'2020, Portland, USA.

[Barbosa et al., 2018] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. (2018). Human mobility: Models and applications. Physics Reports, 734:1 – 74.

[Barrau, 2021] Barrau, Q. (2021). Proposition d'utilisation de la typicité pour aider à la compréhension de clusters de séquences d'activités. Technical report, University of Tours, France.

[Batista et al., 2004] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, 6(1):20–29.

[Becker, 2022] Becker, A. B. (2022). Clustering Methods for Correlated Data. PhD thesis, University of Minnesota, USA.

[Belkin et al., 2003] Belkin, N. J., Kelly, D., Kim, G., Kim, J., Lee, H., Muresan, G., Tang, M. M., Yuan, X., and Cool, C. (2003). Query length in interactive information retrieval. In SIGIR'2003, Toronto, Canada.

[Bellatreche et al., 2005] Bellatreche, L., Giacometti, A., Marcel, P., Mouloudi, H., and Laurent, D. (2005). A personalization framework for OLAP queries. In DOLAP'2005, Bremen, Germany.

[Bie, 2011] Bie, T. D. (2011). An information theoretic framework for data mining. In SIGKDD'2011, San Diego, USA.

[Bie, 2013] Bie, T. D. (2013). Subjective interestingness in exploratory data mining. In IDA'2013, London, UK.

[Bimonte et al., 2023] Bimonte, S., Marcel, P., and Rizzi, S. (2023). Be high on emotion: Coping with emotions and emotional intelligence when querying data. In ADBIS'2023, Barcelona, Spain.

[Bisone, 2021] Bisone, F. (2021). Extraction de trajectoires sémantiques à partir de données multi-capteurs : application à des véhicules de secours. PhD thesis, University of Tours, France.

[Bisone et al., 2019] Bisone, F., Devogele, T., and Etienne, L. (2019). From raw sensor data to semantic trajectories. In EM-GIS'2019 (SIGSPATIAL workshop), Chicago, USA.

[Boniol et al., 2023] Boniol, P., Paparrizos, J., and Palpanas, T. (2023). New trends in time series anomaly detection. In EDBT'2023, Ioannina, Greece.

[Boukharouba et al., 2023] Boukharouba, I., Sèdes, F., Bortolaso, C., and Mouysset, F. (2023). From user activity traces to navigation graph for software enhancement: An application of graph neural network (GNN) on a real-world non-attributed graph. In CIKM'2023, Birmingham, United Kingdom.

[Boukhelifa et al., 2020] Boukhelifa, N., Bezerianos, A., Chang, R., Collins, C., Drucker, S. M., Endert, A., Hullman, J., North, C. L., Sedlmair, M., and Rhyne, T. (2020). Challenges in evaluating interactive visual machine learning systems. IEEE Computer Graphics and Applications, 40(6):88–96.

[Boulil et al., 2014] Boulil, K., Marcel, P., Devogele, T., and Peralta, V. (2014). Projet dopan: des cubes olap pour l'analyse de la vulnérabilité énergétique. In SAGEO'2014 (French conference), Grenoble, France.

[Bres et al., 2022] Bres, R., Peralta, V., Le-Guilcher, A., Devogele, T., Olteanu Raimond, A.-M., and de Runz, C. (2022). Spécification et qualité du réseau cyclable, application à la recherche d'itinéraires. In INFORSID'2022 (French conference), Dijon, France.

[Bres et al., 2023] Bres, R., Peralta, V., Le-Guilcher, A., Devogele, T., Olteanu Raimond, A.-M., and de Runz, C. (2023). Analysis of cycling network evolution in openstreetmap through a data quality prism. In AGILE'2023, Delft, the Netherlands.

[Calegari, 2022] Calegari, D. (2022). Computational narratives using model-driven engineering. In CLEI'2022 (Latin American conference), Quindio, Colombia.

[Carbajal, 2021] Carbajal, S. G. (2021). Customer segmentation through path reconstruction. Sensors, 21(6):2007.

[Carpendale et al., 2016] Carpendale, S., Diakopoulos, N., Riche, N. H., and Hurter, C. (2016). Data-driven storytelling (dagstuhl seminar 16061). Dagstuhl Reports, 6(2):1–27.

[Cauley, 1986] Cauley, K. M. (1986). Studying knowledge acquisition: Distinctions among procedural, conceptual and logical knowledge. In 67th Annual Meeting of the American Educational Research Association.

[Chagnoux, 2020] Chagnoux, M. (2020). La datavisualisation, double point d'entrée du data-journalisme dans la PQR. Interfaces numériques, 9(3).

[Chagnoux et al., 2021] Chagnoux, M., da Silva, R., Outa, F. E., Labroche, N., Marcel, P., and Peralta, V. (2021). Modéliser la démarche du data journaliste : une approche nécessairement transdisciplinaire. In H2PTM'2021 (French conference), Paris, France.

[Chanson et al., 2019] Chanson, A., Crulis, B., Drushku, K., Labroche, N., and Marcel, P. (2019). Profiling user belief in BI exploration for measuring subjective interestingness. In DOLAP'2019, Lisbon, Portugal.

[Chanson et al., 2020] Chanson, A., Crulis, B., Labroche, N., Marcel, P., Peralta, V., Rizzi, S., and Vassiliadis, P. (2020). The traveling analyst problem: Definition and preliminary study. In DOLAP'2020 (EDBT/ICDT workshop), Copenhagen, Denmark.

[Chanson et al., 2021] Chanson, A., Devogele, T., Labroche, N., Marcel, P., Ringuet, N., and T'kindt, V. (2021). A chain composite item recommender for lifelong pathways. In DaWaK'2021, Penang, Malaysia.

[Chanson et al., 2022a] Chanson, A., Labroche, N., Marcel, P., Rizzi, S., and T'kindt, V. (2022a). Automatic generation of comparison notebooks for interactive data exploration. In EDBT'2022, Edinburgh, UK.

[Chanson et al., 2022b] Chanson, A., Outa, F. E., Labroche, N., Marcel, P., Peralta, V., Verdeaux, W., and Jacquemart, L. (2022b). Generating personalized data narrations from EDA notebooks. In DOLAP'2022 (EDBT/ICDT workshop), Edinburgh, UK.

[Chatman, 1980] Chatman, S. (1980). Story and Discourse: Narrative Structure in Fiction and Film. Cornell paperbacks. Cornell University Press.

[Chaudhuri et al., 2011] Chaudhuri, S., Dayal, U., and Narasayya, V. R. (2011). An overview of business intelligence technology. Commun. ACM, 54(8):88–98.

[Chaudhuri and Narasayya, 2007] Chaudhuri, S. and Narasayya, V. R. (2007). Self-tuning database systems: A decade of progress. In VLDB'2007, Vienna, Austria.

[Chédin et al., 2020] Chédin, A., Francia, M., Marcel, P., Peralta, V., and Rizzi, S. (2020). The tell-tale cube. In ADBIS'2020, Lyon, France.

[Chen et al., 2020] Chen, S., Li, J., Andrienko, G. L., Andrienko, N. V., Wang, Y., Nguyen, P. H., and Turkay, C. (2020). Supporting story synthesis: Bridging the gap between visual analytics and storytelling. IEEE Trans. Vis. Comput. Graph., 26(7):2499–2516.

[Cohen-Boulakia, 2022] Cohen-Boulakia, S. (2022). FAIR scientificworkflows: Status, challenges and research opportunities. In EGC'2022, Blois, France.

[Corbett and Anderson, 1995] Corbett, A. T. and Anderson, J. R. (1995). Knowledge tracing: Modelling the acquisition of procedural knowledge. User Model. User Adapt. Interact., 4(4):253–278.

[Crémilleux et al., 2018] Crémilleux, B., Giacometti, A., and Soulet, A. (2018). How your supporters and opponents define your interestingness. In PKDD, Dublin, Ireland.

[Darmont et al., 2022] Darmont, J., Novikov, B., Wrembel, R., and Bellatreche, L. (2022). Advances on data management and information systems. Inf. Syst. Frontiers, 24(1):1–10.

[David et al., 2016] David, Y. B., Segal, A., and Gal, Y. K. (2016). Sequencing educational content in classrooms using bayesian knowledge tracing. In LAK'2016, Edinburgh, UK.

[De Bie et al., 2022] De Bie, T., Raedt, L. D., Hernández-Orallo, J., Hoos, H. H., Smyth, P., and Williams, C. K. I. (2022). Automating data science. Commun. ACM, 65(3):76–87.

[Ding et al., 2019] Ding, R., Han, S., Xu, Y., Zhang, H., and Zhang, D. (2019). Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In SIGMOD'2019, Amsterdam, The Netherlands.

[Djedaini, 2017] Djedaini, M. (2017). Automatic assessment of OLAP exploration quality. PhD thesis, University of Tours, France.

[Djedaini et al., 2019] Djedaini, M., Drushku, K., Labroche, N., Marcel, P., Peralta, V., and Verdeaux, W. (2019). Automatic assessment of interactive OLAP explorations. Inf. Syst., 82:148–163.

[Djedaini et al., 2016] Djedaini, M., Furtado, P., Labroche, N., Marcel, P., and Peralta, V. (2016). Benchmarking exploratory olap. In TPCTC'2016 (VLDB workshop), New Delhi, India.

[Djedaini et al., 2017a] Djedaini, M., Labroche, N., Marcel, P., and Peralta, V. (2017a). A benchmark for assessing OLAP exploration assistants. In EDA'2017 (French conference), Lyon, France.

[Djedaini et al., 2017b] Djedaini, M., Labroche, N., Marcel, P., and Peralta, V. (2017b). Detecting user focus in OLAP analyses. In ADBIS'2017, Nicosia, Cyprus.

[Drushku, 2019] Drushku, K. (2019). User Intent based Recommendation for Modern BI Systems. PhD thesis, University of Tours, France.

[Drushku et al., 2020] Drushku, K., Aligon, J., Labroche, N., Marcel, P., and Peralta, V. (2020). Recommandations basées sur les centres d'intérêts utilisateurs en business intelligence. In INFORSID'2020 (French conference), Dijon, France.

[Drushku et al., 2017] Drushku, K., Aligon, J., Labroche, N., Marcel, P., Peralta, V., and Dumant, B. (2017). User interests clustering in business intelligence interactions. In CAiSE'2017, Essen, Germany.

[Drushku et al., 2019] Drushku, K., Labroche, N., Marcel, P., and Peralta, V. (2019). Interest-based recommendations for business intelligence users. Inf. Syst., 86:79–93.

[Drushku et al., 2021] Drushku, K., Labroche, N., Marcel, P., and Peralta, V. (2021). Learning user interests for recommendations in business intelligence interactions. United States Patent. US 10,915,522 B2.

[Edmond and Bednarz, 2021] Edmond, C. and Bednarz, T. (2021). Three trajectories for narrative visualisation. Vis. Informatics, 5(2):26–40.

[Eichmann et al., 2016] Eichmann, P., Zgraggen, E., Zhao, Z., Binnig, C., and Kraska, T. (2016). Towards a benchmark for interactive data exploration. IEEE Data Eng. Bull., 39(4):50–61.

[Eirinaki et al., 2014] Eirinaki, M., Abraham, S., Polyzotis, N., and Shaikh, N. (2014). Querie: Collaborative database exploration. IEEE Trans. Knowl. Data Eng., 26(7):1778–1790.

[El-Helaly, 2019] El-Helaly, S. (2019). The Mathematics of Voting and Apportionment: An Introduction. Springer International Publishing.

[Elson, 2012] Elson, D. K. (2012). Modeling narrative discourse. PhD thesis, Columbia University, USA.

[Ester et al., 1996] Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD'1996, Portland, USA.

[European Commission et al., 2020] European Commission, Directorate-General for Communications Networks, Content and Technology, Cattaneo, G., Micheletti, G., Glennon, M., La Croce, C., and Mitta, C. (2020). The European data market monitoring tool – Key facts & figures, first policy conclusions, data landscape and quantified stories – D2.9 final study report. Publications Office.

[Förster et al., 2010] Förster, J., Marguc, J., and Gillebaart, M. (2010). Novelty categorization theory. Social and Personality Psychology Compass, 4(9):736 – 755.

[Fradet, 2022] Fradet, V. (2022). L'utilité du data storytelling pour la business intelligence. Technical report, University of Tours, France.

[Francia et al., 2019] Francia, M., Gallinucci, E., and Golfarelli, M. (2019). Social BI to understand the debate on vaccines on the web and social media: unraveling the anti-, free, and pro-vax communities in italy. Soc. Netw. Anal. Min., 9(1):46:1–46:16.

[Francia et al., 2022a] Francia, M., Gallinucci, E., and Golfarelli, M. (2022a). COOL: A framework for conversational OLAP. Inf. Syst., 104:101752.

[Francia et al., 2022b] Francia, M., Gallinucci, E., Golfarelli, M., Marcel, P., Peralta, V., and Rizzi, S. (2022b). Describing multidimensional data through highlights. In SEBD'2022 (Italian conference), Tirrenia, Italy.

[Francia et al., 2023] Francia, M., Golfarelli, M., Marcel, P., Rizzi, S., and Vassiliadis, P. (2023). Suggesting assess queries for interactive analysis of multidimensional data. IEEE Trans. Knowl. Data Eng., 35(6):6421–6434.

[Francia et al., 2022c] Francia, M., Marcel, P., Peralta, V., and Rizzi, S. (2022c). Enhancing cubes with models to describe multidimensional data. Inf. Syst. Frontiers, 24(1):31–48.

[Furtado et al., 2015] Furtado, P., Nadal, S., Peralta, V., Djedaini, M., Labroche, N., and Marcel, P. (2015). Materializing baseline views for deviation detection exploratory OLAP. In DaWaK'2015, Valencia, Spain.

[Gan and Ma, 2023] Gan, M. and Ma, Y. (2023). Mapping user interest into hyper-spherical space: A novel POI recommendation method. Inf. Process. Manag., 60(2):103169.

[Geng and Hamilton, 2006] Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Comput. Surv., 38(3):9.

[Ghodratnama et al., 2021] Ghodratnama, S., Beheshti, A., Zakershahrak, M., and Sobhanmanesh, F. (2021). Intelligent narrative summaries: From indicative to informative summarization. Big Data Res., 26:100257.

[Gkesoulis et al., 2015] Gkesoulis, D., Vassiliadis, P., and Manousis, P. (2015). Cinecubes: Aiding data workers gain insights from OLAP queries. Inf. Syst., 53:60–86.

[Gkitsakis et al., 2022] Gkitsakis, D., Kaloudis, S., Mouselli, E., Peralta, V., Marcel, P., and Vassiliadis, P. (2022). Cube interestingness: Novelty, relevance, peculiarity and surprise. CoRR, abs/2212.03294.

[Gkitsakis et al., 2023] Gkitsakis, D., Kaloudis, S., Mouselli, E., Peralta, V., Marcel, P., and Vassiliadis, P. (2023). Assessment methods for the interestingness of cube queries. In DOLAP'2023 (EDBT/ICDT workshop), Ioannina, Greece.

[Gkitsakis et al., 2024] Gkitsakis, D., Kaloudis, S., Mouselli, E., Peralta, V., Marcel, P., and Vassiliadis, P. (2024). Cube query interestingness: Novelty, relevance, peculiarity and surprise. Inf. Syst., 123:102381.

[Golfarelli and Rizzi, 2009] Golfarelli, M. and Rizzi, S. (2009). Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill.

[Gudivada et al., 2016] Gudivada, V., Irfan, M., Fathi, E., and Rao, D. (2016). Cognitive analytics: Going beyond big data analytics and machine learning. In Gudivada, V. N., Raghavan, V. V., Govindaraju, V., and Rao, C., editors, Cognitive Computing: Theory and Applications, volume 35 of Handbook of Statistics, pages 169–205. Elsevier.

[Guessoum et al., 2022] Guessoum, M. A., Djiroun, R., Boukhalfa, K., and Benkhelifa, E. (2022). Natural language why-question in business intelligence applications: model and recommendation approach. Clust. Comput., 25(6):3875–3898.

[Guha et al., 2015] Guha, R., Gupta, V., Raghunathan, V., and Srikant, R. (2015). User modeling for a personal assistant. In WSDM'2015, Shanghai, China.

[Guidotti et al., 2019] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM Comput. Surv., 51(5):93:1–93:42.

[Gunawardana and Shani, 2009] Gunawardana, A. and Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. J. Mach. Learn. Res., 10:2935–2962.

[Gunawardana and Shani, 2015] Gunawardana, A. and Shani, G. (2015). Evaluating recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, Recommender Systems Handbook, pages 265–308. Springer.

[Gunklach et al., 2023a] Gunklach, J., Jacob, K., and Michalczyk, S. (2023a). Beyond dashboards? designing data stories for effective use in business intelligence and analytics. In ECIS'2023, Kristiansan, Norway.

[Gunklach et al., 2023b] Gunklach, J., Michalczyk, S., Nadj, M., and Maedche, A. (2023b). Metadata extraction from user queries for self-service data lake exploration. Datenbank-Spektrum, 23(2):97–105.

[Gupta and Chandra, 2020] Gupta, M. K. and Chandra, P. (2020). A comprehensive survey of data mining. Int. J. Inf. Technol., 12(4):1243–1257.

[Gyulgyulyan et al., 2019] Gyulgyulyan, E., Aligon, J., Ravat, F., and Astsatryan, H. V. (2019). Data quality alerting model for big data analytics. In QAUCA'2019 (ADBIS workshops), Bled, Slovenia.

[Haddar, 2023] Haddar, I. (2023). Clustering des données pour prédire l'espérance de vie des patients atteints de la sclérose latérale amyotrophique. Master's thesis, National Engineering School of Sfax, Tunisia.

[Hägerstraand, 1970] Hägerstraand, T. (1970). What about people in regional science? Papers in regional science, 24(1):7–24.

[Han et al., 2011] Han, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd edition.

[Harinarayan et al., 1996] Harinarayan, V., Rajaraman, A., and Ullman, J. D. (1996). Implementing data cubes efficiently. In SIGMOD'1996, Montreal, Canada.

[Hawkins et al., 2014] Hawkins, W. J., Heffernan, N. T., and de Baker, R. S. J. (2014). Learning bayesian knowledge tracing parameters with a knowledge heuristic and empirical probabilities. In ITS'2014, Honolulu, USA.

[Herlocker et al., 2004] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. (2004). Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst., 22(1):5–53.

[Hinterberger, 2018] Hinterberger, H. (2018). Exploratory data analysis. In Liu, L. and Özsu, M. T., editors, Encyclopedia of Database Systems. Springer, 2nd edition.

[Huang et al., 2006] Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Schölkopf, B. (2006). Correcting sample selection bias by unlabeled data. In NIPS'2006, Vancouver, Canada.

[Hullman et al., 2013] Hullman, J., Drucker, S. M., Riche, N. H., Lee, B., Fisher, D., and Adar, E. (2013). A deeper understanding of sequence in narrative visualization. IEEE Trans. Vis. Comput. Graph., 19(12):2406–2415.

[Ibrahim et al., 2021] Ibrahim, M. H., Missaoui, R., and Vaillancourt, J. (2021). Detecting important patterns using conceptual relevance interestingness measure. CoRR, abs/2110.11262.

[Idreos et al., 2015] Idreos, S., Papaemmanouil, O., and Chaudhuri, S. (2015). Overview of data exploration techniques. In SIGMOD'2015, Melbourne, Australia.

[Jain et al., 2016] Jain, S., Moritz, D., Halperin, D., Howe, B., and Lazowska, E. (2016). Sqlshare: Results from a multi-year sql-as-a-service experiment. In SIGMOD'2016, San Francisco, USA.

[James and Duncan, 2023] James, S. and Duncan, A. D. (2023). Over 100 data and analytics predictions through 2028. Technical Report G00790199, Gartner, Inc.

[Jayachandran et al., 2014] Jayachandran, P., Tunga, K., Kamat, N., and Nandi, A. (2014). Combining user interaction, speculative query execution and sampling in the DICE system. Proc. VLDB Endow., 7(13):1697–1700.

[Kadastik and Bruni, 2023] Kadastik, N. and Bruni, L. E. (2023). Storifying data for museum audiences. In ExICE'2023, Bologna, Italy.

[Kaminskas and Bridge, 2017] Kaminskas, M. and Bridge, D. (2017). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans. Interact. Intell. Syst., 7(1):2:1–2:42.

[Kaufmann and Rousseeuw, 1987] Kaufmann, L. and Rousseeuw, P. (1987). Clustering by means of medoids. In Statistical Data Analysis based on the L1 Norm, Neuchatel, Switzerland.

[Keim et al., 2018] Keim, D. A., Mansmann, F., Stoffel, A., and Ziegler, H. (2018). Visual analytics. In Liu, L. and Özsu, M. T., editors, Encyclopedia of Database Systems. Springer, 2nd edition.

[Khoussainova et al., 2010] Khoussainova, N., Kwon, Y., Balazinska, M., and Suciu, D. (2010). Snipsuggest: Context-aware autocompletion for SQL. Proc. VLDB Endow., 4(1):22–33.

[Kontarinis et al., 2021] Kontarinis, A., Zeitouni, K., Marinica, C., Vodislav, D., and Kotzinos, D. (2021). Towards a semantic indoor trajectory model: application to museum visits. GeoInformatica, 25(2):311–352.

[Kosara, 2017] Kosara, R. (2017). An argument structure for data stories. In Kozlíková, B., Schreck, T., and Wischgoll, T., editors, EuroVis'2017, Barcelona, Spain.

[Kosara and Mackinlay, 2013] Kosara, R. and Mackinlay, J. D. (2013). Storytelling: The next step for visualization. Computer, 46(5):44–50.

[Kuznetsov et al., 2022] Kuznetsov, S., Tsyryulnikov, A., Kamensky, V., Trachuk, R., Mikhailov, M., Murskiy, S., Koznov, D. V., and Chernishev, G. A. (2022). Unidata - A modern master data management platform. In DataPlat'2022 (workshop of EDBT/ICDT), Edinburgh, UK.

[Kuznetsov and Makhalova, 2018] Kuznetsov, S. O. and Makhalova, T. P. (2018). On interestingness measures of formal concepts. Inf. Sci., 442-443:202–219.

[Lai et al., 2023] Lai, E. Y., Zolaktaf, Z., Milani, M., AlOmeir, O., Cao, J., and Pottinger, R. (2023). Workload-aware query recommendation using deep learning. In EDBT'2023, Ioannina, Greece.

[Lam et al., 2012] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. IEEE Trans. Vis. Comput. Graph., 18(9):1520–1536.

[Lee et al., 2015] Lee, B., Riche, N. H., Isenberg, P., and Carpendale, S. (2015). More than telling a story: Transforming data into visually shared stories. IEEE Computer Graphics and Applications, 35(5):84–90.

[Lemaitre et al., 2017] Lemaitre, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res., 18:17:1–17:5.

[Lesot and Kruse, 2006] Lesot, M. and Kruse, R. (2006). Typicality degrees and fuzzy prototypes for clustering. In GFKL'2006, Berlin, Germany.

[Lezcano Airaldi et al., 2022] Lezcano Airaldi, A., Irrazábal, E., and Diaz-Pace, J. A. (2022). Narrative visualizations best practices and evaluation: A systematic mapping study. PREPRINT (Version 1) available at Research Square, https://doi.org/10.21203/rs.3.rs-1735564/v1. Accessed: 2023-12-27.

[Lipman et al., 2023] Lipman, T., Milo, T., and Somech, A. (2023). ATENA-PRO: generating personalized exploration notebooks with constrained reinforcement learning. In Companion of SIGMOD/PODS'2023, Seattle, WA, USA.

[Litman, 2005] Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. Cognition and Emotion, 19(6):793–814.

[López et al., 2015] López, M. A., Nadal, S., Djedaini, M., Marcel, P., Peralta, V., and Furtado, P. (2015). An approach for alert raising in real-time data warehouses. In EDA'2015 (French conference), Bruxelles, Belgique.

[Ma et al., 2021] Ma, P., Ding, R., Han, S., and Zhang, D. (2021). Metainsight: Automatic discovery of structured knowledge for exploratory data analysis. In SIGMOD'21, pages Xi'an, China.

[Ma et al., 2023] Ma, P., Ding, R., Wang, S., Han, S., and Zhang, D. (2023). Xinsight: explainable data analysis through the lens of causality. Proc. ACM Manag. Data, 1(2):156:1–156:27.

[Marcel et al., 2023a] Marcel, P., Peralta, V., and Amer-Yahia, S. (2023a). Data narration for the people: Challenges and opportunities. In EDBT'2023, tutorial, Ioannina, Greece.

[Marcel et al., 2023b] Marcel, P., Peralta, V., Outa, F. E., and Vassiliadis, P. (2023b). A declarative approach to data narration. CoRR, abs/2303.17141.

[Marcel et al., 2019] Marcel, P., Peralta, V., and Vassiliadis, P. (2019). A framework for learning cell interestingness from cube explorations. In ADBIS'2019, Bled, Slovenia.

[Markl et al., 2018] Markl, V., Borkar, V. R., Zaharia, M., Westmann, T., and Alexandrov, A. (2018). Big data platforms for data analytics. In Liu, L. and Özsu, M. T., editors, Encyclopedia of Database Systems. Springer, 2nd edition.

[McInnes and Healy, 2018] McInnes, L. and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. CoRR, abs/1802.03426.

[Meduri et al., 2021] Meduri, V. V., Chowdhury, K., and Sarwat, M. (2021). Evaluation of machine learning algorithms in predicting the next SQL query from the future. ACM Trans. Database Syst., 46(1):4:1–4:46.

[Megasari et al., 2018] Megasari, M., Wicaksono, P., Li, C. Y., Chaussade, C., Cheng, S., Labroche, N., Marcel, P., and Peralta, V. (2018). Can models learned from a dataset reflect acquisition of procedural knowledge? an experiment with automatic measurement of online review quality. In DOLAP'2018 (EDBT/ICDT workshop), Vienna, Austria.

[Miller, 2017] Miller, H. J. (2017). Time geography. In Shekhar, S., Xiong, H., and Zhou, X., editors, Encyclopedia of GIS, pages 2235–2242. Springer.

[Milo and Somech, 2018] Milo, T. and Somech, A. (2018). Next-step suggestions for modern interactive data analysis platforms. In KDD'2018, London, UK.

[Milo and Somech, 2020] Milo, T. and Somech, A. (2020). Automating exploratory data analysis via machine learning: An overview. In SIGMOD'2020, Portland, USA.

[Mobasher, 2007] Mobasher, B. (2007). Data mining for web personalization. In The Adaptive Web, Methods and Strategies of Web Personalization, volume 4321 of Lecture Notes in Computer Science, pages 90–135. Springer.

[Moreau, 2021] Moreau, C. (2021). Fouille de séquences de mobilité sémantique. PhD thesis, University of Tours, France.

[Moreau et al., 2021a] Moreau, C., Chanson, A., Peralta, V., Devogele, T., and de Runz, C. (2021a). Clustering sequences of multi-dimensional sets of semantic elements. In SAC'2021, Republic of Korea.

[Moreau et al., 2021b] Moreau, C., Devogele, T., de Runz, C., Peralta, V., Moreau, E., and Étienne, L. (2021b). A fuzzy generalisation of the hamming distance for temporal sequences. In FUZZ-IEEE'2021, Luxembourg.

[Moreau et al., 2020a] Moreau, C., Devogele, T., Étienne, L., Peralta, V., and de Runz, C. (2020a). Methodology for mining, discovering and analyzing semantic human mobility behaviors. CoRR, abs/2012.04767.

[Moreau et al., 2020b] Moreau, C., Devogele, T., Peralta, V., and Étienne, L. (2020b). A contextual edit distance for semantic trajectories. In SAC'2020, poster session, Brno, Czech Republic.

[Moreau et al., 2022] Moreau, C., Legroux, C., Peralta, V., and Hamrouni, M. A. (2022). Mining SQL workloads for learning analysis behavior. Inf. Syst., 108:102004.

[Moreau and Peralta, 2021] Moreau, C. and Peralta, V. (2021). Learning analysis behavior in SQL workloads. In DOLAP'2021 (EDBT/ICDT workshop), Nicosia, Cyprus.

[Moreau et al., 2020c] Moreau, C., Peralta, V., Marcel, P., Chanson, A., and Devogele, T. (2020c). Learning analysis patterns using a contextual edit distance. In DOLAP'2020 (EDBT/ICDT workshop), Copenhagen, Denmark.

[Muhammad and Darmont, 2023] Muhammad, F. and Darmont, J. (2023). An ontology-based collaborative business intelligence framework. In DATA'2023, Rome, Italy.

[Nagel et al., 2023] Nagel, D., Affeldt, T., Voges, N., Güntzer, U., and Balke, W. (2023). Binding data narrations - corroborating the plausibility of scientific narratives by open research data. In JCDL'2023, Santa Fe, NM, USA.

[Nejar de Almeida et al., 2023] Nejar de Almeida, V., Ribeiro, E., Bouarour, N., Comba, J. L. D., and Amer-Yahia, S. (2023). SHEVA: A visual analytics system for statistical hypothesis exploration. Proc. VLDB Endow., 16(12):4102–4105.

[Nguyen et al., 2015] Nguyen, H. V., Böhm, K., Becker, F., Goldman, B., Hinkel, G., and Müller, E. (2015). Identifying user interests within the data space - a case study with skyserver. In EDBT'2015, Brussels, Belgium.

[Ondzigue Mbenga, 2023] Ondzigue Mbenga, R. (2023). Système d'information décisionnel, de la narration à la simulation : application à la surveillance épidémiologique de la tuberculose au Gabon. PhD thesis, University of Tours, France, and University of Health Sciences, Gabon.

[Ondzigue Mbenga et al., 2019] Ondzigue Mbenga, R., Devogele, T., Nzondo, S. M., Peralta, V., and Ngoungou, E. B. (2019). Un système d'information géographique décisionnel pour la surveillance épidémiologique de la tuberculose en afrique subsaharienne : Cas du gabon. In SAGEO'2019 (French conference), poster session, Clermont-Ferrand, France.

[Ondzigue Mbenga et al., 2021] Ondzigue Mbenga, R., Peralta, V., Devogele, T., Outa, F. E., Nzondo, S. M., and Ngoungou, E. B. (2021). Processus de narration de données en intelligence épidémique avec application à la pandémie de tuberculose au gabon. In JCIM'2021 (Camerounian conference), Yaoundé, Cameroun.

[Ondzigue Mbenga et al., 2022a] Ondzigue Mbenga, R., Peralta, V., Devogele, T., Outa, F. E., Nzondo, S. M., and Ngoungou, E. B. (2022a). A data narrative about tuberculosis pandemic in gabon. In DARLI-AP'2022 (EDBT/ICDT workshop), Edinburgh, UK.

[Ondzigue Mbenga et al., 2022b] Ondzigue Mbenga, R., Peralta, V., Ngoungou, E. B., Nzondo, S. M., and Devogele, T. (2022b). Narration de données en santé publique: cas de la tuberculose au gabon. In EDA'2022 (French conference), Clermont-Ferrand, France.

[O'Neil et al., 2009] O'Neil, P. E., O'Neil, E. J., Chen, X., and Revilak, S. (2009). The star schema benchmark and augmented fact table indexing. In TPCTC'2009, Lyon, France.

[Oubelmouh et al., 2023] Oubelmouh, Y., Fargon, F., Runz, C. D., Soulet, A., and Veillon, C. (2023). Identifying survival-changing sequential patterns for employee attrition analysis. In DSAA'2023, Thessaloniki, Greece.

[Outa, 2023] Outa, F. E. (2023). A framework for crafting data narratives. PhD thesis, University of Tours, France.

[Outa et al., 2020a] Outa, F. E., Francia, M., Marcel, P., Peralta, V., and Vassiliadis, P. (2020a). Supporting the generation of data narratives. In ER Forum (ER demo section), Vienna, Austria.

[Outa et al., 2020b] Outa, F. E., Francia, M., Marcel, P., Peralta, V., and Vassiliadis, P. (2020b). Towards a conceptual model for data narratives. In ER'2020, Vienna, Austria.

[Outa et al., 2021] Outa, F. E., Marcel, P., and Peralta, V. (2021). Un modèle conceptuel de narration de données. In EDA'2021 (French conference), Toulouse, France.

[Outa et al., 2022] Outa, F. E., Marcel, P., Peralta, V., da Silva, R., Chagnoux, M., and Vassiliadis, P. (2022). Data narrative crafting via a comprehensive and well-founded process. In ADBIS'2022, Turin, Italy.

[Outa et al., 2023] Outa, F. E., Marcel, P., Peralta, V., and Vassiliadis, P. (2023). Highlighting the importance of intentional aspects in data narrative crafting processes. Inf. Syst. Frontiers, 25.

[Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. IEEE Trans. Knowl. Data Eng., 22(10):1345–1359.

[Parent et al., 2013] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G. L., Andrienko, N. V., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., de Macêdo, J. A. F., Pelekis, N., Theodoridis, Y., and Yan, Z. (2013). Semantic trajectories modeling and analysis. ACM Comput. Surv., 45(4):42:1–42:32.

[Parker et al., 2023] Parker, D. C., Sharif, S. V., and Webber, K. (2023). Why did the "missing middle" miss the train? an actors-in-systems exploration of barriers to intensified family housing in waterloo region, canada. Land, 12(2):434.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12:2825–2830.

[Pelánek, 2015] Pelánek, R. (2015). Metrics for evaluation of student models. In EDM'2015, Madrid, Spain.

[Peralta, 2006] Peralta, V. (2006). Data Quality Evaluation in Data Integration Systems. PhD thesis, University of Versailles, France, and University of the Republic, Uruguay.

[Peralta, 2020] Peralta, V. (2020). From source data to data narratives: accompanying users in the way to interactive data analysis. ADBIS/TPDL/EDA'2020 Joint Conferences, keynote, Lyon, France.

[Peralta et al., 2003] Peralta, V., Illarze, A., and Ruggia, R. (2003). On the applicability of rules to automate data warehouse logical design. In DSE'2003 (CAiSE workshop), Klagenfurt/Velden, Austria.

[Peralta et al., 2019a] Peralta, V., Marcel, P., Verdeaux, W., and Diakhaby, A. S. (2019a). Detecting coherent explorations in SQL workloads. CoRR, abs/1907.05618.

[Peralta et al., 2020] Peralta, V., Marcel, P., Verdeaux, W., and Diakhaby, A. S. (2020). Detecting coherent explorations in SQL workloads. Inf. Syst., 92:101479.

[Peralta and Ruggia, 2003] Peralta, V. and Ruggia, R. (2003). Using design guidelines to improve data warehouse logical design. In DMDW'2003 (VLDB workshop), Berlin, Germany.

[Peralta et al., 2019b] Peralta, V., Verdeaux, W., Raimont, Y., and Marcel, P. (2019b). Qualitative analysis of the sqlshareworkload for session segmentation. In DOLAP'2019 (EDBT/ICDT workshop), Lisbon, Portugal.

[Personnaz et al., 2021] Personnaz, A., Amer-Yahia, S., Berti-Équille, L., Fabricius, M., and Subramanian, S. (2021). DORA THE EXPLORER: exploring very large data with interactive deep reinforcement learning. In CIKM'2021, Queensland, Australia.

[Pinon, 2023] Pinon, S. (2023). Business user-oriented recommender system of data. In RCIS'2023, Corfu, Greece.

[Pinon et al., 2022] Pinon, S., Burnay, C., and Linden, I. (2022). Opportunities of semantic recommendation systems for self-service business intelligence. In ICDSST'2022, Thessaloniki, Greece.

[Ranade et al., 2022] Ranade, P., Dey, S., Joshi, A., and Finin, T. (2022). Computational understanding of narratives: A survey. IEEE Access, 10:101575–101594.

[Ranade and Joshi, 2023] Ranade, P. and Joshi, A. (2023). FABULA: intelligence report generation using retrieval-augmented narrative construction. CoRR, abs/2310.13848.

[Rata Gobal, 2022] Rata Gobal, J. (2022). Outils automatiques de narration de données. Technical report, University of Tours, France.

[Ratner et al., 2017] Ratner, A., Bach, S. H., Ehrenberg, H. R., Fries, J. A., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. Proc. VLDB Endow., 11(3):269–282.

[Reisenzein et al., 2012] Reisenzein, R., Meyer, W.-U., and Niepel, M. (2012). Surprise. In Ramachandran, V. S., editor, Encyclopedia of Human Behavior. Elsevier, 2nd edition.

[Ringuet et al., 2022] Ringuet, N., Marcel, P., Labroche, N., Devogele, T., and Bortolaso, C. (2022). Modeling lifelong pathway co-construction. In ER'2022, Hyderabad, India.

[Risam, 2023] Risam, R. (2023). Connecting the dots - refugee data narratives. In Gandhi, E. L. E. and Nguyen, V., editors, The Routledge Handbook of Refugee Narratives. Routledge, 1st edition.

[Rizzi, 2018] Rizzi, S. (2018). Business intelligence. In Liu, L. and Özsu, M. T., editors, Encyclopedia of Database Systems. Springer, 2nd edition.

[Rizzi and Gallinucci, 2014] Rizzi, S. and Gallinucci, E. (2014). Cubeload: A parametric generator of realistic OLAP workloads. In CAiSE'2014, Thessaloniki, Greece.

[Rousseeuw, 1987] Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Computational and Applied Mathematics, 20:53–65.

[Sakka et al., 2021a] Sakka, A., Bimonte, S., Pinet, F., and Sautot, L. (2021a). Volunteer data warehouse: State of the art. Int. J. Data Warehous. Min., 17(3):1–21.

[Sakka et al., 2021b] Sakka, A., Bimonte, S., Rizzi, S., Sautot, L., Pinet, F., Bertolotto, M., Besnard, A., and Rouillier, N. (2021b). A profile-aware methodological framework for collaborative multidimensional modeling. Data Knowl. Eng., 131-132:101875.

[Sapia, 2000] Sapia, C. (2000). Promise: Predicting query behavior to enable predictive caching strategies for olap systems. In DaWaK'2000, London, UK.

[Sarawagi, 2000] Sarawagi, S. (2000). User-adaptive exploration of multidimensional data. In VLDB'2000, Cairo, Egypt.

[Sarawagi et al., 1998] Sarawagi, S., Agrawal, R., and Megiddo, N. (1998). Discovery-driven exploration of OLAP data cubes. In EDBT'1998, Konstanz, Germany.

[Sarker, 2021] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Comput. Sci., 2(3):160.

[Schröder et al., 2023] Schröder, K., Eberhardt, W., Belavadi, P., Ajdadilish, B., van Haften, N., Overes, E., Brouns, T., and Valdez, A. C. (2023). Telling stories with data - A systematic review. CoRR, abs/2312.01164.

[Segel and Heer, 2010] Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data. IEEE Trans. Vis. Comput. Graph., 16(6):1139–1148.

[Serra et al., 2022a] Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2022a). Modeling context for data quality management. In ER'2022, Hyderabad, India.

[Serra et al., 2022b] Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2022b). Use of context in data quality management: a systematic literature review. CoRR, abs/2204.10655.

[Serra et al., 2023] Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2023). Context-aware data quality management methodology. In ADBIS'2023, Barcelona, Spain.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding Machine Learning - From Theory to Algorithms. Cambridge University Press.

[Shi et al., 2021] Shi, D., Xu, X., Sun, F., Shi, Y., and Cao, N. (2021). Calliope: Automatic visual data story generation from a spreadsheet. IEEE Trans. Vis. Comput. Graph., 27(2):453–463.

[Shi, 2022] Shi, Y. (2022). Advances in Big Data Analytics - Theory, Algorithms and Practices. Springer.

[Singh et al., 2007] Singh, V., Gray, J., Thakar, A., Szalay, A. S., Raddick, M. J., Boroski, B., Lebedeva, S., and Yanny, B. (2007). Skyserver traffic report - the first five years. CoRR, abs/cs/0701173.

[Song and Guo, 2016] Song, Y. and Guo, Q. (2016). Query-less: Predicting task repetition for nextgen proactive search and recommendation engines. In WWW'2016, Montreal, Canada.

[Soni et al., 2022] Soni, P., de Runz, C., Bouali, F., and Venturini, G. (2022). Challenges for automatic dashboard generation systems in the context of novice users. In EDA'2022 (French conference), Clermont-Ferrand, France.

[Sugiyama et al., 2007] Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In NIPS'2007, Vancouver, Canada.

[Sun et al., 2023] Sun, M., Cai, L., Cui, W., Wu, Y., Shi, Y., and Cao, N. (2023). Erato: Cooperative data story editing via fact interpolation. IEEE Trans. Vis. Comput. Graph., 29(1):983–993.

[Sun et al., 2016] Sun, Y., Yuan, N. J., Wang, Y., Xie, X., McDonald, K., and Zhang, R. (2016). Contextual intent tracking for personal assistants. In SIGMOD'2016, San Francisco, USA.

[Tejedor et al., 2024] Tejedor, G., Peralta, V., Labroche, N., Marcel, P., Blasco, H., and Alarcan, H. (2024). Stratification pour le pronostic de patients atteints de la sclérose latérale amyotrophique. In EGC'2024 workshop (French conference), Dijon, France.

[Turkey, 1977] Turkey, J. (1977). Exploratory Data Analysis. Addison-Wesley.

[Vashistha and Jain, 2015] Vashistha, A. and Jain, S. (2015). Measuring query complexity in sqlshare workload. Technical report, University of Washington.

[Vassiliadis and Marcel, 2018] Vassiliadis, P. and Marcel, P. (2018). The road to highlights is paved with good intentions: Envisioning a paradigm shift in OLAP modeling. In DOLAP'2018 (EDBT/ICDT workshop), Vienna, Austria.

[Vassiliadis et al., 2024] Vassiliadis, P., Marcel, P., Outa, F. E., Peralta, V., and Gkitsakis, D. (2024). A conceptual model for data storytelling highlights in business intelligence environments. CoRR, abs/2403.00981.

[Vassiliadis et al., 2019] Vassiliadis, P., Marcel, P., and Rizzi, S. (2019). Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. Inf. Syst., 85:68–91.

[Wagner and Fischer, 1974] Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. J. ACM, 21(1):168–173.

[Wang et al., 2013] Wang, H., Song, Y., Chang, M., He, X., White, R. W., and Chu, W. (2013). Learning to extract cross-session search tasks. In WWW'2013, Rio de Janeiro, Brazil.

[Wang and Ma, 2022] Wang, S. and Ma, J. (2022). NMT sentence granularity similarity calculation method based on improved cosine distance. In AIPR'2022, Xiamen, China.

[Wang et al., 2021] Wang, W., Bhowmick, S. S., Li, H., Joty, S. R., Liu, S., and Chen, P. (2021). Towards enhancing database education: Natural language generation meets query execution plans. In SIGMOD'2021, Xi'an, China.

[Wang et al., 2022] Wang, X., Cheng, F., Wang, Y., Xu, K., Long, J., Lu, H., and Qu, H. (2022). Interactive data analysis with next-step natural language query recommendation. CoRR, abs/2201.04868.

[Wang and Heffernan, 2013] Wang, Y. and Heffernan, N. T. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In AIED'2013, Memphis, USA.

[Wang et al., 2020] Wang, Y., Sun, Z., Zhang, H., Cui, W., Xu, K., Ma, X., and Zhang, D. (2020). Datashot: Automatic generation of fact sheets from tabular data. IEEE Trans. Vis. Comput. Graph., 26(1):895–905.

[Wang et al., 2019] Wang, Z., Wang, S., Farinella, M., Murray-Rust, D., Riche, N. H., and Bach, B. (2019). Comparing effectiveness and engagement of data comics and infographics. In CHI'2019, Glasgow, Scotland, UK.

[Watson, 2006] Watson, R. T. (2006). Data management - databases and organizations (5. ed.). Wiley.

[White and Roth, 2009] White, R. W. and Roth, R. A. (2009). Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool Publishers.

[Wong et al., 2006] Wong, M., Bhattarai, B., and Singh, R. (2006). Characterization and analysis of usage patterns in large multimedia websites. Technical report, San Francisco State University, USA.

[Wongsuphasawat et al., 2019] Wongsuphasawat, K., Liu, Y., and Heer, J. (2019). Goals, process, and challenges of exploratory data analysis: An interview study. CoRR, abs/1911.00568.

[Yao et al., 2006] Yao, Y., Chen, Y., and Yang, X. D. (2006). A measurement-theoretic foundation of rule interestingness evaluation. In Foundations and Novel Approaches in Data Mining. Springer Berlin Heidelberg.

[Youngmann et al., 2022] Youngmann, B., Amer-Yahia, S., and Personnaz, A. (2022). Guided exploration of data summaries. Proc. VLDB Endow., 15(9):1798–1807.

[Yuan et al., 2021] Yuan, J., Chen, C., Yang, W., Liu, M., Xia, J., and Liu, S. (2021). A survey of visual analytics techniques for machine learning. Comput. Vis. Media, 7(1):3–36.

[Zhang and Sun, 2023] Zhang, N. and Sun, S. (2023). Multiview unsupervised shapelet learning for multivariate time series clustering. IEEE Trans. Pattern Anal. Mach. Intell., 45(4):4981–4996.

[Zolaktaf et al., 2020] Zolaktaf, Z., Milani, M., and Pottinger, R. (2020). Facilitating SQL query composition and analysis. In SIGMOD'2020, Portland, USA.

[Zreik, 2023] Zreik, A. (2023). Semantic trajectory analysis for the prediction of the physical state of the collections at the BnF. PhD thesis, University of Paris-Saclay, France.

[Zuo et al., 2019] Zuo, J., Zeitouni, K., and Taher, Y. (2019). Incremental and adaptive feature exploration over time series stream. In BigData, Los Angeles, USA.

Appendices

Appendix A

Studied query workloads

In what follows, we consider query workloads, typically arising from logs of database systems or query tools, containing (potentially long) sequences of queries made by some users. In particular, we focus on workloads of hand-written[1] queries.

Most of the workloads arise from experiments specially designed to test an analysis tool or platform, described in the literature [Aligon et al., 2014a, Jain et al., 2016, Drushku et al., 2017, Milo and Somech, 2018]. Another workload results from the testing phase of a research project [Boulil et al., 2014]. They all consist of navigation traces of real users on real data. Unlike them, some additional workloads were artificially generated using specialized generation tools [Rizzi and Gallinucci, 2014].

In our experiments, we reuse such workloads for new tasks, as evaluating the quality of explorations, learning users' behavior and skills and discovering users' interests.

We chose to test our proposals in several workloads to avoid learning specific behavior of a set of users. Indeed, the considered workloads concern users with different analysis skills (students, novices, experts), using different analysis tools (open source tools, research prototypes, advanced user interfaces) and accessing datasets of different sizes and complexities. We are not aware of other public analytical workloads, specially from senior analysts, whose analysis activity is jealously guarded by companies as pointed out by [Rizzi and Gallinucci, 2014].

In the following sections we describe, for each workload, its origins, users and underlying data. We also report a quantitative description of the workloads in terms of sessions, explorations and queries, and describe additional metadata if available. Finally, we describe a user study and the workload devised for supporting it.

A.1 Workloads of real users' queries

Ipums. The first workload, henceforth dubbed *Ipums*, consists of navigation traces of OLAP users, collected in 2014, during the testing phase of the development of Falseto [Aligon et al., 2014a], a tool meant to assist query and exploration composition, by letting the user summarize, browse, query, and reuse former analytical explorations. The 17 OLAP users engaged in the test were students of two Master's programs specialized in Business Intelligence. The test was not part of the programs, was not graded and all the participants were volunteers. They developed

[1] Consistently with the authors of [Jain et al., 2016], we use the term hand-written to mean, in this context, that the query is introduced manually by a human user, which reflects genuine interactive human activity over a dataset, with consideration between two consecutive queries.

explorations for answering four analytical questions on the IPUMS cube, for example, *"Are energy costs following the evolution of the average income for some profiles?"* and *"Where is it better to live in terms of energy costs, for an individual profile?"*.

The IPUMS cube integrates data from the IPUMS (Integrated Public Use Microdata Series) website[2]. It is organized as a star schema with 5 dimensions (year, city, sex, race and occupation), 12 (non-top) levels, 25 measures, and contains 500,000 facts recorded in the fact table.

From this experiment, we reuse 27 explorations counting 306 queries, with an average of 11 queries per exploration.

Some metadata is also exploited, as the workload also logs the users devising the queries and the analytical questions being answered. In addition, for their original experiment (i.e. evaluating Falseto tool), Aligon et al. labelled explorations distinguishing five analysis styles [Aligon et al., 2014a].

- **FOCUS**. The exploration is more focused as time passes,
- **OSCILLATE-FOCUS**. The exploration is more exploratory (the levels of detail and filtering oscillate) at the beginning but is more focused at the end,
- **OSCILLATE**. The exploration is always exploratory,
- **FIX**. The exploration keeps constant levels of detail and filtering,
- **ATYPICAL**. The exploration has atypical or erratic behavior.

We reuse such labels as kind of ground truth for our experiments on users' analysis behavior.

Open. The second workload, henceforth dubbed *Open*, consists of navigation traces collected in 2016, in the context of the DOPAn project on energy vulnerability (described in Appendix C).

The underlying dataset contains 3 data cubes. In the main cube, called **MobPro**, facts represent people trips between home and workplace, and dimensions allow to characterize a trip according to various characteristics of the worker (e.g. age, gender, level of studies), home (e.g. location, family size), job (e.g. location, branch), transport mode, traveled distance, energy used, etc. The main cube is organized as a star schema with 19 dimensions, 68 (non-top) levels, 24 measures, and contains 37,149 facts recorded in the fact table. The other cubes are organized in a similar way.

Navigation traces were produced by 10 volunteer students of a Master's degree in Business Intelligence, answering fuzzy information needs defined by their lecturer. Students were asked to explore data in order to gain insights about mobility profiles and energy consumption. However, students were not aware that navigation traces could be used for research tasks, not to perturbed their behavior and bias experiments. During their task, students investigated some relations among data, for example, *"Which profiles of workers, having low revenues, expend the most in mobility"* and tested some popular hypothesis like *"Executives make longer home-work trips than people with other professions"*.

To explore the cube, the students used Saiku[3] a web application that allows to navigate OLAP databases in a user-friendly manner, and generates MDX code from graphical manipulations. The students were quite familiar with this OLAP tool, but not necessarily with the data in the cube.

From this experiment, we could gather 39 explorations from the system logs. In total, these explorations represent 1608 queries, with an average of 41 queries per exploration. A particularity of some third party OLAP tools, like Saiku, is that their user interfaces submit a new query for

[2]Minnesota Population Center. Integrated Public Use Microdata Series. `http://www.ipums.org`
[3]Saiku OLAP tool: `http://meteorite.bi/products/saiku`

each user action (including intermediate drag-and-drops), resulting in very long explorations in the log. Nevertheless, there were some extremely short explorations, which mainly correspond to incomplete studies.

Additional metadata includes users and query timestamps. Explorations were also manually inspected by the lecturer and tagged with A-B-C labels according to their overall quality. Label A corresponds to good explorations, clearly following an information need, investigating it and containing coherent queries. Students producing such explorations are considered to have analysis skills. Contrarily, label C denotes poor explorations, with less contributory queries, typically switching topics, with no clear information need. Label B corresponds to students that are learning analysis skills, but still produce middle-quality explorations.

In order to build a ground truth about query quality, queries were labeled by experts, using a labeling tool specifically designed for that purpose. Queries were independently annotated by two experts (lecturers) for learning focus (with a high agreement of 89%) and two additional experts (interns working on OLAP exploration) for learning query contribution (we keep the queries where both agreed, 67,59%).

Enterprise. The third workload, henceforth dubbed *Enterprise*, consists in navigation traces of 14 volunteers of SAP company[4] in the context of a research and innovation project [Drushku, 2019], recorded in 2017.

Analysts covered a range of skills in data exploration, classed, based on their position in the company, in two expertise groups: beginners and expert users. They were asked to analyze some of the 7 available data sources to answer 10 business needs (named Q_1 to Q_{10}), each corresponding to a specific user interest. The business needs were grouped in different business cases like: *"For each European country, detect which genres of films did not reach the expected sales"* or *"In which Income Group would you classify a candidate country with a GDP of $6 billion?"*.

Table A.1 describes, for each business need, its difficulty, estimated by an expert (in terms of time, number of queries and exploited sources expected in its resolving), the number of sessions and queries devised for solving it, and the number of queries perceived as relevant by users in their own activity.

	Business needs									
	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
Difficulty	low	med	med	med	low	high	low	low	med	high
Number of sessions	19	11	10	10	10	8	9	9	9	8
Number of queries	84	65	60	41	50	43	61	51	26	49
Number of relevant queries	34	26	30	16	26	10	27	24	24	9
Queries / session	4.4	5.9	6.0	4.1	5.0	5.4	6.8	5.7	2.9	6.1
Relevant queries / session	1.8	2.4	3.0	1.6	2.6	1.25	3.0	2.7	2.7	1.1

Table A.1: Analysis of business needs

They used a SAP prototype that supports keyword-based BI queries[5].

As users enter keywords, the BI system suggests, on the fly, further keywords to complete the current ones, letting the user choose among them, as in web search engines. The underlying idea is that a suggestion completes the original BI question in order to obtain a well-formed query over a database.

[4]SAP company web site: `https://www.sap.com/about.html`

[5]Patent Reference: 14/856,984 : BI Query and Answering using full text search and keyword semantics

In total, this workload contains 24 sessions, accounting for 105 explorations and 530 queries. We remark that conversely to Saiku tool, SAP prototype only evaluates final queries, after all keywords were entered and a formal query was selected. In this context, average exploration length is around 5 queries.

The prototype log is very rich. It includes users, entered keywords, suggestions (additional keywords and related database queries), the chosen suggestions and the result of executing them. There are no timestamps. Knowledge on users' skills is also available.

Queries and explorations were also tagged according to their quality, as done for the Open workload. For the sake of confidentiality, they were labeled by only one expert (5 queries, considered as outliers, were not labeled).

Security. The fourth workload, henceforth dubed *Security*, consists of analysis sessions made by real analysts in the context of the Honeynet Project[6]. 56 analysts specialized in the domain of cyber-security were recruited (via dedicated forums, network security firms, and volunteer senior students from the Israeli National Cyber-Security Program) and were asked to analyze 4 different datasets using a prototype web-based analysis platform developed for the project [Milo and Somech, 2018].

Each dataset contains between 350 to 13K rows of raw network logs that may reveal a distinct security event, e.g. malware communication hidden in network traffic, hacking activity inside a local network, an IP range/port scan, etc. (there is no connection between the tuples of different datasets). The analysts were asked to perform as many analysis actions as required to reveal the details of the underlying security event of each dataset.

From this workload, we reuse 723 explorations and 3868 queries, with an average of 5 queries per exploration. It is particularly interesting because queries were devised by expert analysts.

As additional metadata, the workload contains user ids, project id (referring to the 4 datasets) and timestamps.

SQLShare. The SQLShare workload is the result of a Multi-Year SQL-as-a-Service Experiment [Jain et al., 2016], allowing any user with minimal database experience to upload their datasets on-line and manipulate them via SQL queries. What the authors wanted to prove with this experiment is that SQL is beneficial for data scientists. They observed that most of the time people use scripts to modify or visualize their datasets instead of using the SQL paradigm. Indeed, most user needs may be satisfied by first-order queries, that are much simpler than a script, but have the initial cost of creating a schema, importing the data and so on. SQL-as-a-Service frees the user of all this prior work with a relaxed SQL version.

The SQLShare workload is composed of 11,137 SQL statements (of which, 10,668 are queries), 97 users and 3,336 user's datasets. To the best of our knowledge and as reported by the authors of [Jain et al., 2016], this workload is the only one containing primarily ad-hoc hand-written queries over user-uploaded datasets.

The SQLShare workload is analyzed in [Jain et al., 2016], particularly to verify the following assumption: *"We hypothesized that SQLShare users would write queries that are more complex individually and more diverse as a set, making the corpus more useful for designing new systems."*. The authors showed empirically that the queries in the SQLShare workload are complex and diverse. They also analyzed the churn rate of SQLShare users and conclude that most users exhibit a behavior that suggest an exploratory workload.

[6]Honeynet Project: `https://www.honeynet.org/`

Characteristics	Ipums	Open	Enterprise	Security	SQLShare
Nb sessions		25	24		451
Nb explorations	27	39	104	723	
Nb queries	306	1608	530	3868	10,668
Nb users	17	10	14	56	97
User expertise	students	students	beginners & experts	experts	unknown
Analysis tool	Falseto	Saiku	SAP tool	web UI	SQL UI
Dataset size	500K facts	$\sim$ 100K facts	not reported	$\sim$ 30K rows	varied

Table A.2: Summary of workload characteristics

Table A.2 summarizes the main characteristics of the 5 workloads. Notably, in Ipums, Open, Enterprise and Security workloads, users did not have to write any SQL code, contrarily to SQLShare. Indeed, the used analytical tools generated queries from users' high-level operations. However, in both cases, users devised real explorations, taking the time to analyse results before devising new queries. Users of the Ipums and Open workloads were Master students learning data analysis, users of the Enterprise workload were developers with varied analysis skills, users of the Security workload were expert analysts, while SQLShare users are anonymous end-users and there is no knowledge about their analysis skills.

A.2 Synthetic workloads

Artificial. The Artificial workload, consists of artificial explorations devised over artificial data from the Star Schema Benchmark (SSB) [O'Neil et al., 2009]. SSB is a variation of TPC-H, a popular benchmark from the Transaction Processing Performance Council (TPC).

SSB cube consists of a relational database under the form of a star schema, with one fact table and 4 dimension tables.

Instead of using the rather limited SSB workload, we generated artificial explorations using CubeLoad [Rizzi and Gallinucci, 2014], a tool for generating realistic explorations over star schemas. CubeLoad takes as input a cube schema and creates the desired number of explorations according to templates modeling various user exploration patterns.

CubeLoad proposes four templates that simulate recurrent types of user analyses, namely:

- **Slice And Drill**. Following the default behavior of several OLAP front-ends, hierarchies are progressively navigated by choosing a member of a current group-by level, creating a selection predicate on such member and drilling down on it. Therefore, explorations of this template contain sequences of filter and drill-down operations.
- **Slice All**. Users are sometimes interested in navigating a cube by slices, i.e., repeatedly running the same query but with different selection predicates. Then, this templates generates sequences of unfilter/filter operations.
- **Exploratory**. The motivation for this template is the assumption that several users, while exploring the cube in search of significant correlations, will be "attracted" by one surprising query and then evolve casually. So, explorations based on this template contain varied random operations.
- **Goal Oriented**. Explorations of this type are run by users who have a specific analysis goal, but whose OLAP skills are limited so they may follow a complex path to reach their destination. Explorations of this template contain varied operations but converging to some specific point.

From this workload, we reuse 50 explorations and 908 queries, with an average of 18 queries per exploration.

Loan. The second synthetic workload, henceforth dubed *Loan*, consists of some explorations over artificial data from the Loan cube of the PKDD99 Discovery Challenge[7].

The instances of the cube were generated with a dedicated random generator, producing 3 versions of increasing size, specifically: 100,000, 1 million, and 10 million facts.

This workload is used for scalability tests; the semantics of explorations is not exploited.

A.3 User study

We also conducted a user study[8] in order to evaluate how do the interestingness aspects (those presented in Chapter 3) relate to the behavior of people working with cubes and cube queries. This section describes the user study and the underlying workload.

Adult. The last workload, henceforth dubed *Adult*, consists of few explorations devised by researchers, each query carefully chosen to maximize an interestingness aspect. Notice that even if the explorations are devised by humans, they do not represent genuine analysis activities, as queries are assembled artificially for the user study.

The Adult dataset[9] is used, as it is very easy to understand. The cube contains census data, organized in 8 dimensions (*Age, Native Country, Education, Occupation, Marital Status, Work Class, Gender and Race*) and a single measure, *Work Hours Per Week*.

25 students from France and Greece participated to the study. There were 7 PhD and 11 Master students, all trained in Business Intelligence concepts, and 7 undergraduate students with significantly less exposure to such concepts.

Each participant received a short description of the Adult dataset structure and semantics, and the overall business goal of *finding out which are the categories of working people with significantly higher and lower average working hours per week*. They were given some pre-computed queries along with their results to help them determine the answer to the task. All queries in the experiment were expressed in natural language and their results were presented as crosstabs. Specifically, participants received a warming-up set of query results that give a broad description of how working hours are related to various cube dimensions. Subsequently, participants received 3 sets of 4 queries to be ranked according to their interest (from 1=most interesting to 4=less interesting, without ties) and were asked to justify their ranks.

[7] PKDD99 Discovery Challenge: `https://sorry.vse.cz/~berka/challenge/pkdd1999/chall.htm`

[8] All the material of the study, along with our findings, are available via a public repository: `https://github.com/OLAP3/2023InterestingnessUserStudy`.

[9] Adult dataset: `https://archive.ics.uci.edu/dataset/2/adult`

Appendix B

Studied data narratives

In what follows, we consider several data narratives, either crafted during user studies[1], or publicly available and described by their authors.

B.1 Challenges

Narrating Rennes by the data. We organized a one-day challenge[2] during a workshop and we observed several narrators with various profiles while they crafted data narratives for answering the challenge. Their goal was to produce data narratives using the open data of the French city of Rennes[3].

Three teams (A, B, C) were constituted, mixing one or two data enthusiasts (among which journalists, students, social workers) and a data scientist. An external observer (lecturer or PhD student in Computer Science) annotated the crafting process followed by each team. In particular, they wrote down the sequences of activities that were performed.

It should be noted that the teams were allowed to continue their crafting work during 3 additional days. During the annotation period (only the initial day, during the workshop), all teams mainly performed exploration and question answering activities; only one team (C) started the structuring and presentation of the data narrative. Importantly, the teams were not asked to follow the process we propose; only the observers were aware of it. The produced data narratives -a video, a notebook and an interactive book- (in French) are publicly available[1]. A prize was awarded to the best one.

Fatal encounters. We organized a second challenge for Master students, and asked an experienced data journalist to assess the quality of data narratives crafted during the challenge, and the completion of each phase of the process.

Participants were 44 Master students in Computer Science, specialized in data analysis, 14 of the first year of master (hereafter called M1), 30 of the second year (called M2). Obviously, M2 students have more experience with data analysis and visualization tools, however, all students were familiar with the dataset (they previously did some data cleaning tasks in class) and none of them had previous experience with data narratives.

[1]The data narratives crafted during challenges are available at:
`https://drive.google.com/drive/folders/1zDzP_ndSlQUJCbtFMVzJDnIbyXK1D2_l?usp=sharing`

[2]Challenge Narrating Rennes by the data, sponsored by MaDICS and CNRS:
`https://www.madics.fr/event/titre1617704707-3351/#madona`.

[3]Open data of Rennes: `https://data.rennesmetropole.fr/`

Students were asked to craft a data narrative about fatal encounters in the USA, using an open dataset[4]. They received a one-hour tutorial on data narratives, presenting definitions and examples, and introducing typical crafting activities. Students worked by pairs or alone. We received 7 data narratives from M1 students and 17 from M2 students.

B.2 Real data narratives

Figures B.1-B.5 show (part of) five data narratives with different visual styles, namely, an infographics, a scrolly-story, a news article, a news article with interactive graphics, and a video.

They are described below:

- **Strokes.** A data narrative informing women about stroke risks, in the form of an infographic[5]. It combines different visualizations (circles, bars, percentages), with pictograms and text for providing explanations. Highlighted measures renforce the messages.
  The data narrative was crafted by Justin McKinley, computer graphics designer, and published by GOOD, a B-corp social-impact company[6].
- **Climate.** A data narrative about the climate crisis in the Sahel, in the form of a scrolly-story[7]. As we scroll in the web page that tells the story, several links allow to explore the underlying data. For instance, the link "Explore displacement data →" leads to some interactive graphics explaining displacement issues (as the two included at the bottom of Figure B.2). The reader can do some navigation and apply filters to obtain more details.
  The data narrative was crafted by Julia Janicki, data journalist, and published by OCHA[8], the United Nations Office for the Coordination of Humanitarian Affairs.
  The crafting process is documented in the blog of the data journalist[9], describing her analytical questions and hypotheses, and providing examples of queries and findings. The structuring and presentation is also described.
- **Tennis.** A data narrative about racket in tennis betting, in the form of sport news[10]. It is mainly textual, with many numbers reinforcing the discourse.
  The data narrative was crafted by Heidi Blake, investigations editor, and John Templon, investigative data reporter, and published by BuzzFeed News[11].
  The crafting process is documented by the data reporter, also in the form of sport news[12], detailing his analytical questions and the corresponding data exploration.
- **Covid.** A data narrative about covid excess mortality in Alsace in the form of a news article with interactive visualizations[13] (in French). The bottoms near the graphics (as the "Bas-Rhin" and "Haut-Rhin" ones in the screenshot) allow interaction. Highlighting individual curves are also possible.

[4]Fatal encounters: `https://fatalencounters.org/`

[5]Facts About Women and Strokes: `https://www.good.is/infographics/facts-about-women-and-strokes`

[6]GOOD: `https://www.good.is/`

[7]The Climate Crisis in the Sahel: `https://data.humdata.org/visualization/climate-crisis-sahel/`

[8]OCHA: `https://www.unocha.org/ocha`

[9]Developing a data story on the climate crisis in the Sahel:
`https://centre.humdata.org/developing-a-data-story-on-the-climate-crisis-in-the-sahel/`

[10]The Tennis Racket:
`https://www.buzzfeednews.com/article/heidiblake/the-tennis-racket#.nnZ8bYLw2`

[11]FuzzFeed News: https://www.buzzfeednews.com/

[12]How BuzzFeed News Used Betting Data To Investigate Match-Fixing In Tennis: `https://www.buzzfeednews.com/article/johntemplon/how-we-used-data-to-investigate-match-fixing-in-tennis`

[13]Toutes causes confondues, la Covid a tué jusqu'à cinq fois plus d'Alsaciens pendant la crise:
`https://tinyurl.com/24ubaanu`

The data narrative was crafted by Raphaël da Silva, data journalist, and published by Rue89 Strasbourg Newspaper[14].
The crafting process is documented by the data journalist in a notebook[15]. His analytical questions are detailed, accompanied with python code for analysing the data. In a local copy of the notebook, it is possible to modify the code and continue the investigation with new queries.

– **Tuberculosis**. A data narrative about Tuberculosis pandemic in Gabon, in the form of a Business Intelligence application (for internal use of the Gabonese Ministry of Health) and published as a video[16] (In French). The video shows a particular route throught the interactive application, recording interactive manipulation of the dashboards, with oral explanations of the main findings. Messages are highlighted all along.
The data narrative was developed by Raymond Ondzigue Mbenga, statistics manager of the eGabon-SIS project[17]. The crafting process is documented in a research article [Ondzigue Mbenga et al., 2022a]. All steps are described, with examples.

[14]Rue89 Strasbourg: `https://www.rue89strasbourg.com/`
[15]Raphël Da Silva's blog: `https://tinyurl.com/yc5chu57`
[16]`https://www.youtube.com/watch?v=u_KoBWc_qJU`
[17]eGabon-SIS project - `https://www.facebook.com/p/Projet-Egabon-SIS-100064718617026/`

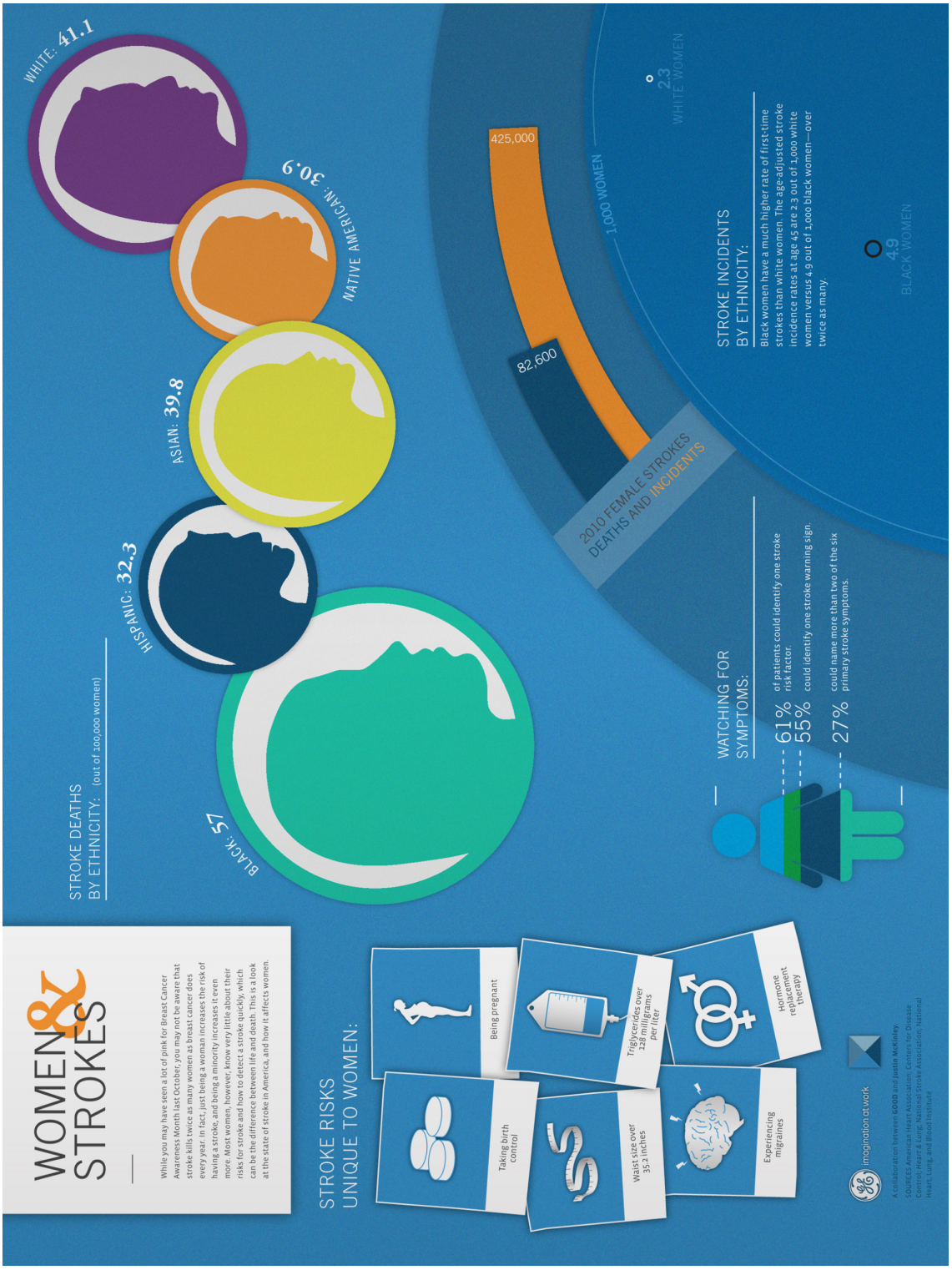


Figure B.1: Facts About Women and Strokes (by GOOD)

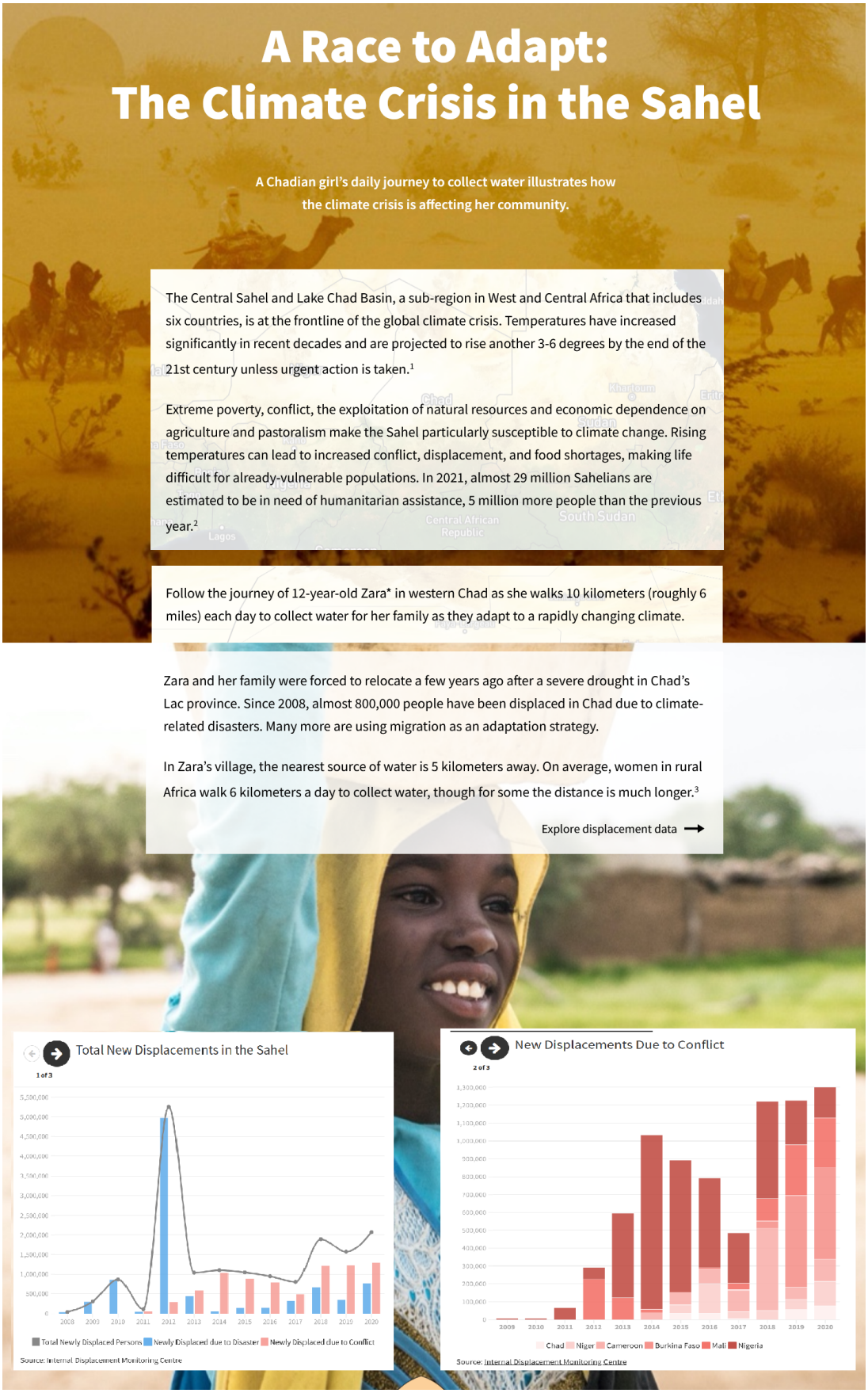


Figure B.2: (Some screenshot of) The Climate Crisis in the Sahel (by OCHA)

WORLD • THE TENNIS RACKET

The Tennis Racket

Betting worth billions. Elite players. Violent threats. Covert messages with Sicilian gamblers. And suspicious matches at Wimbledon. Leaked files expose match-fixing evidence that tennis authorities have kept secret for years.

Heidi Blake
UK Investigations Editor, UK

John Templon
BuzzFeed News Reporter

Posted on January 17, 2016 at 10:58 pm

View 12 comments

Secret files exposing evidence of widespread match-fixing by players at the upper level of world tennis can today be revealed by BuzzFeed News and the BBC.

The sport's governing bodies have been warned repeatedly about a core group of 16 players – all of whom have ranked in the top 50 – but none have faced any sanctions and more than half of them will begin playing at the Australian Open on Monday.

It has been seven years since world tennis authorities were first handed compelling evidence about a network of players suspected of fixing matches at major tournaments including Wimbledon following a landmark investigation, but all of them have been allowed to continue playing.

The investigation into men's tennis by BuzzFeed News and the BBC is based on a cache of leaked documents from inside the sport – the Fixing Files – as well as an original analysis of the betting activity on 26,000 matches and interviews across three continents with gambling and match-fixing experts, tennis officials, and players.

The files contain detailed evidence of suspected match-fixing orchestrated by gambling syndicates in Russia and Italy, which was uncovered in the landmark 2008 probe, and which authorities subsequently shelved. "They could have got rid of a network of players that would have almost completely cleared the sport up," said Mark Phillips, one of the investigators. "We gave them everything tied up with a nice pink bow on top and they took no action at all."

BuzzFeed News began its investigation after devising an algorithm to analyse gambling on professional tennis matches over the past seven years. It identified 15 players who regularly lost matches in which heavily lopsided betting appeared to substantially shift the odds – a red flag for possible match-fixing.

Four players showed particularly unusual patterns, losing almost all of these red-flag matches. Given the bookmakers' initial odds, the chances that the players would perform that badly were less than 1 in 1,000. (Read more about the analysis here.)

Tennis is the latest sport to be caught up in allegations of corruption following the scandals that have engulfed world football and athletics.

It can today be revealed:

- Winners of singles and doubles titles at Grand Slam tournaments are among the core group of 16 players who have repeatedly been reported for losing games when highly suspicious bets have been placed against them.
- One top-50 player competing in the Australian Open is suspected of repeatedly fixing his first set.

Figure B.3: (The initial part of) The Tennis Racket (by BuzzFeed News)

Rue89 Strasbourg
ENQUÊTES ET ACTUALITÉ À STRASBOURG ET EUROMÉTROPOLE

Connexion S'abonner

Toutes causes confondues, la Covid a tué jusqu'à cinq fois plus d'Alsaciens pendant la crise

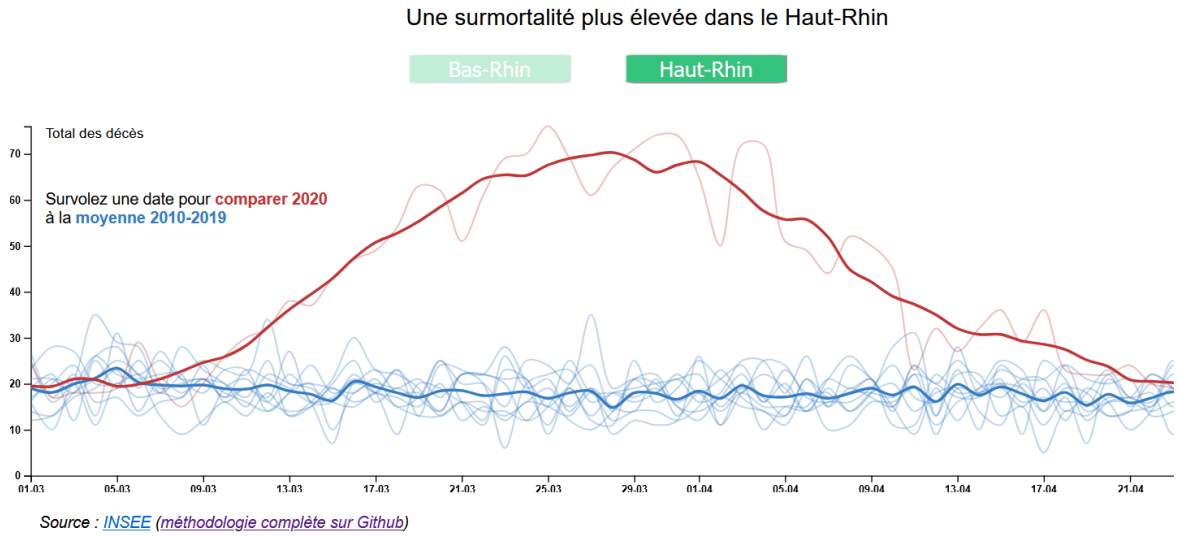
Rue89 Strasbourg a produit une visualisation de la mortalité de mars et d'avril dans les départements alsaciens, en la comparant à la décennie précédente. Le doyen de la Faculté de médecine de Strasbourg et le chef des urgences de Colmar partagent leur analyse.

+ #Covid-19 44

Cet article est en accès libre. Pour soutenir Rue89 Strasbourg, abonnez-vous.

Raphaël da Silva

Publié le 18 juin 2020 · 4 minutes



Source : INSEE (méthodologie complète sur Github)

Depuis la mi-mars, Santé Publique France et l'Agence régionale de santé (ARS) ont quotidiennement actualisé les chiffres des hospitalisations, des patients en réanimation, des retours à domicile… et des décès. Mais cette dernière donnée n'était qu'une estimation, basée sur les remontées (parfois incomplètes) des établissements hospitaliers et des Ehpad.

En utilisant les chiffres de l'Insee, il est possible d'afficher une visualisation plus exhaustive de la surmortalité départementale liée à la covid-19.

Figure B.4: (The initial part of) All causes combined, Covid killed up to five times more Alsatians during the crisis (by Rue89 Strasbourg)

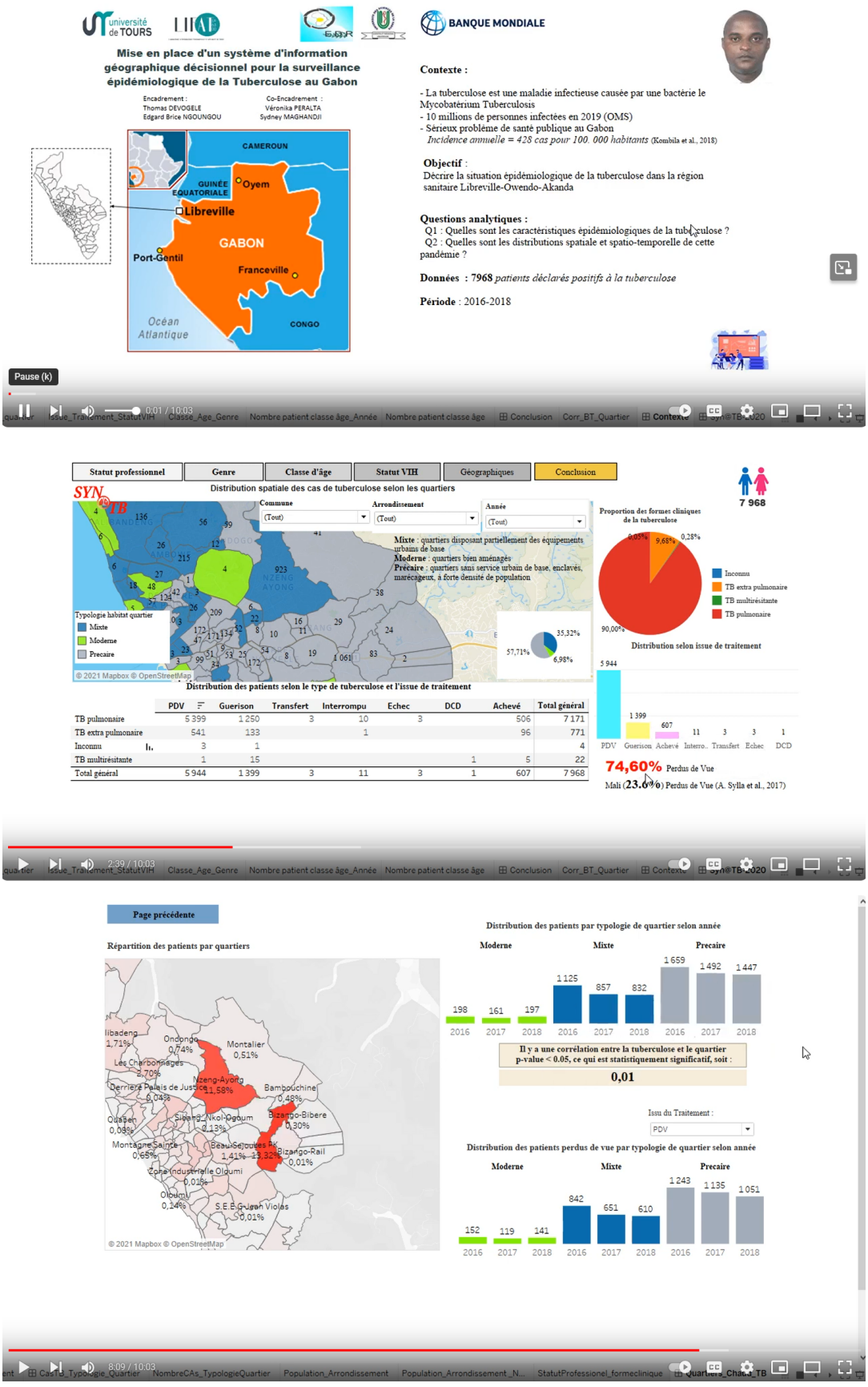


Figure B.5: (Some screenshots of) Epidemic intelligence of Tuberculosis in Gabon (by eGabonSIS)

Appendix C

Research projects

This appendix describes several projects related to the topics addressed in this dissertation. See Appendix E for a complete list of projects.

DOPAn - Open Data for Monitoring and Analysis[1] (2014-2017)

Description: The DOPAn project focuses on the interactive analysis of open data, to study the energy vulnerability of the households and territories of Loir et Cher French department, regarding domestic energy and mobility-related expenses. The outcome of the project is a user-friendly, user-centered dedicated Business Intelligence solution, extensible to other types of data and territories. The project aims to: (i) Build a data warehouse to centralize data and prepare analysis, collecting primary information that is not available yet. (ii) Support analysts and study their analytical habits to adapt analysis tools to their needs. (iii) Develop a dedicated solution, easy to use, durable, and extended to other types of data and other territories.

Consortium: LIFAT - University of Tours; Observatory of Economy and Territories of Loir et Cher. Leader: Patrick Marcel (LIFAT)

Funding: French regional funding

Web site: https://lifat.univ-tours.fr/lifat-english-version/projects/recent/bdtin/2014-2017-dopan

[1]Original name (in French): **DOPAn** - **D**onnées **O**uvertes pour le **P**ilotage et l'**An**alyse

Mobi'Kids - The role of urban educative cultures in the evolution of children's daily mobility and life context. Collection and analysis of geolocated and semantically enriched tracks[2] (2017-2021)

Description: The MOBI'KIDS program aims to understand the conditions of children's daily mobility and their relationship to the city spaces in a context of changing lifestyles and encouraged alternative modes of travel spurred by the "city demotorization". The goal is also to study how children's mobility and spatial experiences evolve during the transition from elementary school to secondary school. It is about going further the notion of routine to detect more informal or spontaneous forms of children's practices of space. The comparison of different living contexts (city center vs. urban periphery) contributes to the characterization of "urban educational cultures" (UEC) as well as the drivers of behavioral changes and thus to support public policies. Producing detailed and localized knowledge of family mobilities and practices in urban spaces involves designing a mixed data collection and analysis protocol that combines quantitative and qualitative approaches based on methods and paradigms from different disciplines. It relies on two industrial partners for the deployment of optimized sensors, filtering and sequencing algorithms for geolocalized data, and an online information system to enrich and qualify travel and activity trace data. The proposed protocol provides knowledge that is complementary to that obtained from standard mobility surveys (Household travel survey).

Consortium: UMR AAU; LIFAT - University of Tours; ALKANTE; RF TRACK; UMR ESO; PACTE - University of Grenoble Alpes. Leader: Sandrine Depeau (ESO)

Funding: National funding (ANR)

Web site: https://anr.fr/Projet-ANR-16-CE22-0009

Madona - Mastering Interactive Data Analysis for Journalistic Narration[3] (2018-2022)

Description: The project aims to bring together computer science researchers, specialists in interactive data exploration, and researchers in information and communication science, specialists in digital production and reception practices, in interaction with journalists with data mining issues (data journalism). The objectives of these meetings are: (i) a better understanding of the mechanisms of data selection and exploration which make it possible to gradually construct data narratives: by understanding the empirical methodologies adopted by journalists and by questioning the scripting of data journalism productions; (ii) the creation and provision of tools allowing journalists to explore and analyze open data with simplified interaction: by formulating hypotheses and expressing needs with high-level primitives, by selecting or refining proposals of models or visualizations extracted from data.

Consortium: IRIT - University of Toulouse; CREM - University of Rennes; LIFAT - University of Tours; Rue 89 Strasbourg. Leaders: Julien Aligon (IRIT), Marie Chagnoux (CREM), Patrick Marcel (LIFAT)

Funding: French national funding (MaDICS, CNRS)

Web site: https://sites.google.com/view/action-madics-madona/home

[2]Original name (in French): **Mobi'Kids** - Le rôle des cultures éducatives urbaines dans l'évolution des mobilités quotidiennes et des contextes de vie des enfants. Collecte et analyse de traces géolocalisées et enrichies sémantiquement

[3]Original name (in French): **Madona** – **M**aîtriser l'**A**nalyse interactive de **DO**nnées pour la **NA**rration journalistique

OPTIMEDIAS - Optimization of Data Exploitation by Artificial Intelligence in Health[4] (2022-2025)

Description: The acquisition of massive health data raises questions about their structuring and exploitation to optimize medicine precision. While the use of Artificial Intelligence appears essential to respond to this problem, the observation of efficient but dispersed regional wealth motivates the establishment of a collaborative project around this theme. This project aims to bring together academic and industrial stakeholders to promote health data through Artificial Intelligence in order to develop algorithms for predicting diagnosis, prognosis and therapeutic decision support.

Consortium: CHRU of Tours, iBrain, CEPR and LIFAT - University of Tours; ATOS; Orleans Val de Loire Technopole; LIFO - University of Orleans. Leader: Hélène Blasco (CHRU)

Funding: French regional funding

JUNON - Digital twins for natural resources[5] (2022-2027)

Description: At a time of digital and environmental transition, the Centre-Val de Loire Region and its partners aimed to create a single innovative digital hub dedicated to the environment as well as the management of natural resources. The high point of this approach is the development of digital twins on water, soil and air using the latest developments in Artificial Intelligence. Digital twins make it possible to reproduce real structures and processes throughout their evolution using Machine Learning approaches. These virtual reproductions are built around learning that requires large quantities of data and is based on expertise in environmental metrology. JUNON brings together around twenty research teams distributed across ten collaborative projects working together to: (i) create and exploit new digital twins, (ii) establish new public and commercial services to manage our natural resources more sustainably.

Consortium: BRGM, Centre-Val de Loire Region; University of Orleans; University of Tours; INRAE; CNRS; Orleans Val de Loire Technopole; DREAM Water & Environment Pole; AgreenTech Valley; Nextino; FarmViz; ANTEA group; ATOS; SDEC; Le Studium; AgroPithiviers; Lig'Air; Le Gabor 45; Orleans Métropole. Leader: Sébastien Dupraz (BRGM)

Funding: French regional funding (ARD)

Web site: https://www.brgm.fr/fr/programme/junon-jumeaux-numeriques-au-service-ressources-naturelles

Data quality within data preparation for Big Data analysis[6] (2023-2026)

Description: This project aims is to solve data quality management issues in architectures for big data analysis. The considered architectures are those that combine Data Lake and Data Warehouse on the same platform. In these architectures, the data has a variety of formats and degrees of structuring, and may be in different stages of processing, even in different storage areas within the platform. These areas are used by different types of users for different types of analysis. A Context-based approach will be followed for the solution. This approach constitutes a very important and current line of research in the area of Data Quality.

Consortium: InCo - University of the Republic (Uruguay); LIFAT - University of Tours. Leader: Adriana Marotta (InCo)

Funding: Uruguayan national funding (CSIC)

[4]Original name (in French): **OPTIMEDIAS** - **OPTIM**isation de l'**E**xploitation des **D**onnées par l'**I**ntelligence *A*rtificielle en **S**anté

[5]Original name (in French): **JUNON** – Des jumeaux numériques au service des ressources naturelles

[6]Original name (in Spanish): Calidad de Datos en la Preparación para el Análisis de Big Data

IntForOut - Multisource spatial data **Int**egration **For** the monitoring of ecosystems under the pressure of **Out**door recreation (2024-2027)

Description: Upheavals in practices (speed of adaptation of human mobility, increasing use of natural spaces for recreational purposes) have a strong impact on the ecosystem (flora and fauna). The IntForOut project aims to develop a methodological framework to integrate heterogeneous and fragmentary data on human activities, biodiversity and ecosystems, in order to: (i) improve the assessment of human pressure on ecosystems, (ii) design solutions to achieve a compromise between outdoor activities and nature conservation, and (iii) promote new uses of data produced by stakeholders.

Consortium: LASTIG - University Gustave Eiffel/IGN/ENSG; LIFAT - University of Tours; LECA & OFB LIG - University of Savoie Mont-Blanc 2; CREA Mont-Blanc. Leader: Ana Maria Olteanu Raimond

Funding: French national funding (ANR)

Appendix D

List of publications

International Journals

1. Gkitsakis, D., Kaloudis, S., Mouselli, E., Peralta, V., Marcel, P., and Vassiliadis, P. (2024). Cube query interestingness: Novelty, relevance, peculiarity and surprise. *Inf. Syst.*, 123:102381
2. Outa, F. E., Marcel, P., Peralta, V., and Vassiliadis, P. (2023). Highlighting the importance of intentional aspects in data narrative crafting processes. *Inf. Syst. Frontiers*, 25
3. Moreau, C., Legroux, C., Peralta, V., and Hamrouni, M. A. (2022). Mining SQL workloads for learning analysis behavior. *Inf. Syst.*, 108:102004
4. Francia, M., Marcel, P., Peralta, V., and Rizzi, S. (2022b). Enhancing cubes with models to describe multidimensional data. *Inf. Syst. Frontiers*, 24(1):31–48
5. Peralta, V., Marcel, P., Verdeaux, W., and Diakhaby, A. S. (2020). Detecting coherent explorations in SQL workloads. *Inf. Syst.*, 92:101479
6. Djedaini, M., Drushku, K., Labroche, N., Marcel, P., Peralta, V., and Verdeaux, W. (2019). Automatic assessment of interactive OLAP explorations. *Inf. Syst.*, 82:148–163
7. Drushku, K., Labroche, N., Marcel, P., and Peralta, V. (2019). Interest-based recommendations for business intelligence users. *Inf. Syst.*, 86:79–93
8. Berti-Équille, L., Comyn-Wattiau, I., Cosquer, M., Kedad, Z., Nugier, S., Peralta, V., Cherfi, S. S., and Thion-Goasdoué, V. (2011). Assessment and analysis of information quality: a multidimensional model and case studies. *Int. J. Inf. Qual.*, 2(4):300–323

National Journals

9. Peralta, V. and Bouzeghoub, M. (2006). Data freshness evaluation in different application scenarios. *Revue des Nouvelles Technologies de l'information (French journal)*, E5:373–378
10. Peralta, V., Ruggia, R., and Bouzeghoub, M. (2004a). Analyzing and evaluating data freshness in data integration systems. *Ingénierie des Systèmes d'Information (French journal)*, 9(5-6):145–162

Proceedings

11. Aligon, J. and Peralta, V., editors (2021). *Business Intelligence & Big Data, Proceedings of EDA'2021 (French conference)*, volume B-17 of *RNTI*, Toulouse, France. Editions RNTI

Book chapters and theses

12. Marotta, A., Cancela, H., Peralta, V., and Ruggia, R. (2010). Reliability models for data integration systems. In Faulin, J., Juan, A. A., Martorell, S., and Ramírez-Márquez, J.-E., editors, *Simulation Methods for Reliability and Availability of Complex Systems*, volume XVIII of *Springer Series on Reliability Engineering*, pages 123–144. Springer
13. Peralta, V. (2006b). *Data Quality Evaluation in Data Integration Systems*. PhD thesis, University of Versailles, France, and University of the Republic, Uruguay
14. Peralta, V. (2001). Diseño lógico de data warehouses a partir de esquemas conceptuales multidimensionales. Master's thesis, University of the Republic, Montevideo, Uruguay

International Conferences

15. Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2023). Context-aware data quality management methodology. In *ADBIS'2023*, Barcelona, Spain
16. Gkitsakis, D., Kaloudis, S., Mouselli, E., Peralta, V., Marcel, P., and Vassiliadis, P. (2023). Assessment methods for the interestingness of cube queries. In *DOLAP'2023 (EDBT/ICDT workshop)*, Ioannina, Greece
17. Bres, R., Peralta, V., Le-Guilcher, A., Devogele, T., Olteanu Raimond, A.-M., and de Runz, C. (2023). Analysis of cycling network evolution in openstreetmap through a data quality prism. In *AGILE'2023*, Delft, the Netherlands
18. Outa, F. E., Marcel, P., Peralta, V., da Silva, R., Chagnoux, M., and Vassiliadis, P. (2022). Data narrative crafting via a comprehensive and well-founded process. In *ADBIS'2022*, Turin, Italy
19. Chanson, A., Outa, F. E., Labroche, N., Marcel, P., Peralta, V., Verdeaux, W., and Jacquemart, L. (2022). Generating personalized data narrations from EDA notebooks. In *DOLAP'2022 (EDBT/ICDT workshop)*, Edinburgh, UK
20. Ondzigue Mbenga, R., Peralta, V., Devogele, T., Outa, F. E., Nzondo, S. M., and Ngoungou, E. B. (2022a). A data narrative about tuberculosis pandemic in gabon. In *DARLI-AP'2022 (EDBT/ICDT workshop)*, Edinburgh, UK
21. Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2022a). Modeling context for data quality management. In *ER'2022*, Hyderabad, India
22. Moreau, C. and Peralta, V. (2021). Learning analysis behavior in SQL workloads. In *DOLAP'2021 (EDBT/ICDT workshop)*, Nicosia, Cyprus
23. Moreau, C., Devogele, T., de Runz, C., Peralta, V., Moreau, E., and Étienne, L. (2021b). A fuzzy generalisation of the hamming distance for temporal sequences. In *FUZZ-IEEE'2021*, Luxembourg
24. Moreau, C., Chanson, A., Peralta, V., Devogele, T., and de Runz, C. (2021a). Clustering sequences of multi-dimensional sets of semantic elements. In *SAC'2021*, Republic of Korea
25. Chédin, A., Francia, M., Marcel, P., Peralta, V., and Rizzi, S. (2020). The tell-tale cube. In *ADBIS'2020*, Lyon, France
26. Chanson, A., Crulis, B., Labroche, N., Marcel, P., Peralta, V., Rizzi, S., and Vassiliadis, P. (2020). The traveling analyst problem: Definition and preliminary study. In *DOLAP'2020 (EDBT/ICDT workshop)*, Copenhagen, Denmark
27. Moreau, C., Peralta, V., Marcel, P., Chanson, A., and Devogele, T. (2020c). Learning analysis patterns using a contextual edit distance. In *DOLAP'2020 (EDBT/ICDT workshop)*, Copenhagen, Denmark

28. Outa, F. E., Francia, M., Marcel, P., Peralta, V., and Vassiliadis, P. (2020b). Towards a conceptual model for data narratives. In *ER'2020*, Vienna, Austria
29. Peralta, V., Verdeaux, W., Raimont, Y., and Marcel, P. (2019b). Qualitative analysis of the sqlshareworkload for session segmentation. In *DOLAP'2019 (EDBT/ICDT workshop)*, Lisbon, Portugal
30. Marcel, P., Peralta, V., and Vassiliadis, P. (2019). A framework for learning cell interestingness from cube explorations. In *ADBIS'2019*, Bled, Slovenia
31. Megasari, M., Wicaksono, P., Li, C. Y., Chaussade, C., Cheng, S., Labroche, N., Marcel, P., and Peralta, V. (2018). Can models learned from a dataset reflect acquisition of procedural knowledge? an experiment with automatic measurement of online review quality. In *DOLAP'2018 (EDBT/ICDT workshop)*, Vienna, Austria
32. Djedaini, M., Labroche, N., Marcel, P., and Peralta, V. (2017b). Detecting user focus in OLAP analyses. In *ADBIS'2017*, Nicosia, Cyprus
33. Drushku, K., Aligon, J., Labroche, N., Marcel, P., Peralta, V., and Dumant, B. (2017). User interests clustering in business intelligence interactions. In *CAiSE'2017*, Essen, Germany
34. Djedaini, M., Furtado, P., Labroche, N., Marcel, P., and Peralta, V. (2016). Benchmarking exploratory olap. In *TPCTC'2016 (VLDB workshop)*, New Delhi, India
35. Furtado, P., Nadal, S., Peralta, V., Djedaini, M., Labroche, N., and Marcel, P. (2015). Materializing baseline views for deviation detection exploratory OLAP. In *DaWaK'2015*, Valencia, Spain
36. Ba, C., da Costa, U. S., Ferrari, M. H., Ferré, R., Musicante, M. A., Peralta, V., and Robert, S. (2014). Preference-driven refinement of service compositions. In *CLOSER'2014*, Barcelona, Spain
37. Aligon, J., Boulil, K., Marcel, P., and Peralta, V. (2014). A holistic approach to olap sessions composition: The falseto experience. In *DOLAP'2014 (CIKM workshop)*, Shanghai, China
38. Romero, O., Marcel, P., Abelló, A., Peralta, V., and Bellatreche, L. (2011). Describing analytical sessions using a multidimensional algebra. In *DaWaK'2011*, Toulouse, France
39. González, L., Peralta, V., Bouzeghoub, M., and Ruggia, R. (2009). Qbox-services: Towards a service-oriented quality platform. In *QoIs'2009 (ER workshop)*, Gramado, Brazil
40. Peralta, V., Thion-Goasdoué, V., Kedad, Z., Berti-Équille, L., Comyn-Wattiau, I., Nugier, S., and Cherfi, S. S. (2009b). Multidimensional management and analysis of quality measures for CRM applications in an electricity company. In *ICIQ'2009*, Potsdam, Germany
41. Peralta, V., Kostadinov, D., and Bouzeghoub, M. (2009a). Apmd-workbench: A benchmark for query personalization. In *CIRSE'2009 (ECIR workshop)*, Toulouse, France
42. Akoka, J., Berti-Équille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué-Thion, V., Kedad, Z., Nugier, S., Peralta, V., and Cherfi, S. S. (2007). A framework for quality evaluation in data integration systems. In *ICEIS'2007*, Funchal, Portugal
43. Grigori, D., Peralta, V., and Bouzeghoub, M. (2005). Service retrieval based on behavioral specifications and quality requirements. In *BPM'2005*, Nancy, France
44. Bouzeghoub, M. and Peralta, V. (2004). A framework for analysis of data freshness. In *IQIS'2004 (SIGMOD workshop)*, Paris, France
45. Peralta, V., Illarze, A., and Ruggia, R. (2003a). On the applicability of rules to automate data warehouse logical design. In *DSE'2003 (CAiSE workshop)*, Klagenfurt/Velden, Austria

46. Peralta, V. and Ruggia, R. (2003). Using design guidelines to improve data warehouse logical design. In *DMDW'2003 (VLDB workshop)*, Berlin, Germany
47. Peralta, V. (2003a). Data warehouse logical design from multidimensional conceptual schemas. In *CLEI'2003*, La Paz, Bolivia
48. Peralta, V. and Marotta, A. (2002). Hacia la automatización del diseño de data warehouses. In *CLEI'2002*, Montevideo, Uruguay

National Conferences

49. Tejedor, G., Peralta, V., Labroche, N., Marcel, P., Blasco, H., and Alarcan, H. (2024). Stratification pour le pronostic de patients atteints de la sclérose latérale amyotrophique. In *EGC'2024 workshop (French conference)*, Dijon, France
50. Ondzigue Mbenga, R., Peralta, V., Ngoungou, E. B., Nzondo, S. M., and Devogele, T. (2022b). Narration de données en santé publique: cas de la tuberculose au gabon. In *EDA'2022 (French conference)*, Clermont-Ferrand, France
51. Francia, M., Gallinucci, E., Golfarelli, M., Marcel, P., Peralta, V., and Rizzi, S. (2022a). Describing multidimensional data through highlights. In *SEBD'2022 (Italian conference)*, Tirrenia, Italy
52. Bres, R., Peralta, V., Le-Guilcher, A., Devogele, T., Olteanu Raimond, A.-M., and de Runz, C. (2022). Spécification et qualité du réseau cyclable, application à la recherche d'itinéraires. In *INFORSID'2022 (French conference)*, Dijon, France
53. Outa, F. E., Marcel, P., and Peralta, V. (2021). Un modèle conceptuel de narration de données. In *EDA'2021 (French conference)*, Toulouse, France
54. Ondzigue Mbenga, R., Peralta, V., Devogele, T., Outa, F. E., Nzondo, S. M., and Ngoungou, E. B. (2021). Processus de narration de données en intelligence épidémique avec application à la pandémie de tuberculose au gabon. In *JCIM'2021 (Camerounian conference)*, Yaoundé, Cameroun
55. Chagnoux, M., da Silva, R., Outa, F. E., Labroche, N., Marcel, P., and Peralta, V. (2021). Modéliser la démarche du data journaliste : une approche nécessairement transdisciplinaire. In *H2PTM'2021 (French conference)*, Paris, France
56. Drushku, K., Aligon, J., Labroche, N., Marcel, P., and Peralta, V. (2020). Recommandations basées sur les centres d'intérêts utilisateurs en business intelligence. In *INFORSID'2020 (French conference)*, Dijon, France
57. Djedaini, M., Labroche, N., Marcel, P., and Peralta, V. (2017a). A benchmark for assessing OLAP exploration assistants. In *EDA'2017 (French conference)*, Lyon, France
58. López, M. A., Nadal, S., Djedaini, M., Marcel, P., Peralta, V., and Furtado, P. (2015). An approach for alert raising in real-time data warehouses. In *EDA'2015 (French conference)*, Bruxelles, Belgique
59. Boulil, K., Marcel, P., Devogele, T., and Peralta, V. (2014). Projet dopan: des cubes olap pour l'analyse de la vulnérabilité énergétique. In *SAGEO'2014 (French conference)*, Grenoble, France
60. Etcheverry, L., Peralta, V., and Bouzeghoub, M. (2008). Qbox-foundation: a metadata platform for quality measurement. In *DKQ'2008 (workshop of EGC, French conference)*, Sophia-Antipolis, France
61. Akoka, J., Berti-Équille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué-Thion, V., Kedad, Z., Nugier, S., Peralta, V., Quafafou, M., and Cherfi, S. S. (2008). Évaluation de la qualité des systèmes multisources : Une approche par les patterns. In *DKQ'2008 (workshop of EGC, French conference)*, Sophia-Antipolis, France

62. Sastre, D., Peralta, V., and Ruggia, R. (2008). Evaluación de calidad en una aplicacioón de data warehousing: de la definición de metas a la especificación de métricas. In *WBD'2008 (Chilean workshop)*, Punta Arenas, Chile
63. Bouzeghoub, M., Calabretto, S., Denos, N., Harrathi, R., Kostadinov, D., Nguyen, A., and Peralta, V. (2007). Accès personnalisé aux informations: approche dirigée par la qualité. In *INFORSID'2007 (French conference)*, Perros-Guirec, France
64. Bouzeghoub, M. and Peralta, V. (2005). Data freshness evaluation in different application scenarios. In *DKQ'2005 (workshop of EGC, French conference)*, Paris, France
65. Peralta, V. and Bouzeghoub, M. (2004). On the evaluation of data freshness in data integration systems. In *BDA'2004 (French conference)*, Montpellier, France
66. Peralta, V., Ruggia, R., Kedad, Z., and Bouzeghoub, M. (2004b). A framework for data quality evaluation in a data integration system. In *SBBD'2004, (Braziian conference)*, Brasília, Brazil
67. Fajardo, F., Crsipino, I., and Peralta, V. (2004). Dqe: Una herramienta para evaluar la calidad de los datos en unsistema de integración. In *CACIC'2004, (Argentinian conference)*, La Matanza, Argentina
68. Motz, R., Ruggia, R., Abin, J., Marotta, A., Carpani, F., and Peralta, V. (2003). Proyecto SICO: sistemas de información en un entorno cooperativo. In *JISBD'2003 (Spanish conference)*, Alicante, Spain
69. Peralta, V. (2002). Un escenario para diseño lógico de data warehouses. In *JCC'2002 (Chilean workshop)*, Copiapo, Chile
70. Peralta, V. and Kedad, Z. (2002). Una plataforma basada en metadata para cálculo de vistas en sistemas de data warehousing. In *CACIC'2002 (Argentinian conference)*, Buenos Aires, Argentina

Posters and demos

71. Outa, F. E., Francia, M., Marcel, P., Peralta, V., and Vassiliadis, P. (2020a). Supporting the generation of data narratives. In *ER Forum (ER demo section)*, Vienna, Austria
72. Moreau, C., Devogele, T., Peralta, V., and Étienne, L. (2020b). A contextual edit distance for semantic trajectories. In *SAC'2020, poster session*, Brno, Czech Republic
73. Ondzigue Mbenga, R., Devogele, T., Nzondo, S. M., Peralta, V., and Ngoungou, E. B. (2019). Un système d'information géographique décisionnel pour la surveillance épidémiologique de la tuberculose en afrique subsaharienne : Cas du gabon. In *SAGEO'2019 (French conference), poster session*, Clermont-Ferrand, France
74. Kostadinov, D., Peralta, V., Soukane, A., and Xue, X. (2005). In *INFORSID (French conference), demo session*, Grenoble, France
75. Kostadinov, D., Peralta, V., Soukane, A., and Xue, X. (2004). Système adaptatif d'aide à la génération de requêtes de médiation. In *BDA'2004 (French conference), demo session*, Montpellier, France
76. Peralta, V., Illarze, A., and Ruggia, R. (2003b). Towards the automation of data warehouse logical design: a rule-based approach. In *CAiSE Forum'2003 (CAiSE demo)*, Klagenfurt/Velden
77. Peralta, V., Marotta, A., and Ruggia, R. (1999). Designing data warehouses through schema transformation primitives. In *ER'1999 (poster and demo)*, Paris, France

Patents

78. Drushku, K., Labroche, N., Marcel, P., and Peralta, V. (2021). Learning user interests for recommendations in business intelligence interactions. United States Patent. US 10,915,522 B2

Keynotes and tutorials

79. Marcel, P., Peralta, V., and Amer-Yahia, S. (2023a). Data narration for the people: Challenges and opportunities. In *EDBT'2023, tutorial*, Ioannina, Greece
80. Marcel, P. and Peralta, V. (2022a). Data exploration from insights to storytelling. In *eBISS'2022 (Summer School), tutorial*, Cesena, Italy
81. Marcel, P. and Peralta, V. (2022b). Exploratory data analysis: from insights to storytelling. In *é-EGC'2022 (Winter School of EGC, French conference) on "Humans in the data exploration and learning loop", tutorial*, Blois, France
82. Peralta, V. (2020). From source data to data narratives: accompanying users in the way to interactive data analysis. ADBIS/TPDL/EDA'2020 Joint Conferences, keynote, Lyon, France

Preprints and technical reports

83. Vassiliadis, P., Marcel, P., Outa, F. E., Peralta, V., and Gkitsakis, D. (2024). A conceptual model for data storytelling highlights in business intelligence environments. *CoRR*, abs/2403.00981
84. Marcel, P., Peralta, V., Outa, F. E., and Vassiliadis, P. (2023b). A declarative approach to data narration. *CoRR*, abs/2303.17141
85. Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2022b). Use of context in data quality management: a systematic literature review. *CoRR*, abs/2204.10655
86. Gkitsakis, D., Kaloudis, S., Mouselli, E., Peralta, V., Marcel, P., and Vassiliadis, P. (2022). Cube interestingness: Novelty, relevance, peculiarity and surprise. *CoRR*, abs/2212.03294
87. Moreau, C., Devogele, T., Étienne, L., Peralta, V., and de Runz, C. (2020a). Methodology for mining, discovering and analyzing semantic human mobility behaviors. *CoRR*, abs/2012.04767
88. Peralta, V., Marcel, P., Verdeaux, W., and Diakhaby, A. S. (2019a). Detecting coherent explorations in SQL workloads. *CoRR*, abs/1907.05618
89. Peralta, V. (2007). Extraction and integration of movielens and imdb data. Technical report, University of Versailles, Versailles, France
90. Peralta, V. (2006a). Data freshness and data accuracy: A state of the art. Technical report, University of the Republic, Montevideo, Uruguay
91. Peralta, V., Marotta, A., and Ruggia, R. (2003c). Towards the automation of data warehouse design. Technical report, University of the Republic, Montevideo, Uruguay
92. Peralta, V. (2003b). Data warehouse logical design from multidimensional conceptual schemas. Technical report, University of the Republic, Montevideo, Uruguay

Appendix E

Curriculum Vitæ

This appendix lists the main activities of my Curriculum Vitæ, except for publications, as they were listed in Appendix D.

Verónika Peralta

Associate Professor – Head of CS department

LIFAT Laboratory, University of Tours

1 Identity

Uruguayan and French.	Born in August 13th 1974, married, 2 children.
Work address:	Université de Tours – Antenne de Blois 3 place Jean Jaurès, 41000 Blois, France Tél: +33 (0) 2 54 55 21 12
E-mail:	veronika.peralta@univ-tours.fr
Web page:	http://www.info.univ-tours.fr/~veronika/

2 Education

2003-2006	**PhD in Computer Science**, University of Versailles (France) and University of the Republic (Uruguay). Dissertation: Data quality evaluation in data integration systems
1998-2001	**Master in Computer Science**, University of the Republic (Uruguay). Dissertation: Data Warehouse Logical Design from Multidimensional Conceptual Schemas
1993-1998	**Engineer Degree**, University of the Republic (Uruguay). Dissertation: Study of techniques and tools for the development of data warehousing systems

3 Positions

From 2008	Associate Professor, University of Tours (France)
2005-2008	Associate Professor, University of the Republic (Uruguay)
2008	Lecturer, University of Buenos Aires (Argentina)
2006-2007	Research Engineer (postdoc), University of Versailles (France)
2006-2007	Lecturer, University of Versailles (France)
1996-2005	Lecturer, University of the Republic (Uruguay)
1996-2003	Lecturer, South Autonomous University (Uruguay)
1998-2001	Computer Science Engineer, Associate Consulting Engineers (consulting firm, Uruguay)
1997-1998	Programming Analyst, Associate Consulting Engineers (consulting firm, Uruguay)

4 Research activities

My recent research activities address the general challenge of developing techniques to support exploratory data analysis, from data collection to the visualization of a data story. I am particularly interested in personalization, interactivity and automation aspects. More precisely, my activities are focused on learning the interests, intentions and habits of the user and taking advantage to guide analysis, automate steps and recommend actions adapted to each user profile. Particular interest is given to the quality of an interactive analysis.

My research topics include: data analysis, quality of analysis, data quality, data stories, personalization, recommendation and OLAP, particularly in the context of heterogeneous, autonomous and distributed information systems, especially in big data context, dealing with poorly (or not) structured data, of varied format and quality.

Selected publications:

Hereafter is a selection of publications (see the complete list at the end) and a quantitative summary by publication type and scope. My h-index, as calculated by Google Scholar, is currently 16.

1. Outa, F. E., Marcel, P., Peralta, V., and Vassiliadis, P. (2023). Highlighting the importance of intentional aspects in data narrative crafting processes. Inf. Syst. Frontiers, 25
2. Serra, F., Peralta, V., Marotta, A., and Marcel, P. (2023). Context-aware data quality management methodology. In ADBIS'2023, Barcelona, Spain
3. Moreau, C., Legroux, C., Peralta, V., and Hamrouni, M. A. (2022). Mining SQL workloads for learning analysis behavior. Inf. Syst., 108:102004
4. Francia, M., Marcel, P., Peralta, V., and Rizzi, S. (2022b). Enhancing cubes with models to describe multidimensional data. Inf. Syst. Frontiers, 24(1):31–48
5. Moreau, C., Devogele, T., de Runz, C., Peralta, V., Moreau, E., and Etienne, L. (2021b). A fuzzy generalisation of the hamming distance for temporal sequences. In FUZZ-IEEE'2021, Luxembourg
6. Djedaini, M., Drushku, K., Labroche, N., Marcel, P., Peralta, V., and Verdeaux, W. (2019). Automatic assessment of interactive OLAP explorations. Inf. Syst., 82:148–163
7. Drushku, K., Labroche, N., Marcel, P., and Peralta, V. (2019). Interest-based recommendations for business intelligence users. Inf. Syst., 86:79–93
8. Marcel, P., Peralta, V., and Vassiliadis, P. (2019). A framework for learning cell interestingness from cube explorations. In ADBIS'2019, Bled, Slovenia
9. Berti-Equille, L., Comyn-Wattiau, I., Cosquer, M., Kedad, Z., Nugier, S., Peralta, V., Cherfi, S. S., and Thion-Goasdoué, V. (2011). Assessment and analysis of information quality: a multidimensional model and case studies. Int. J. Inf. Qual., 2(4):300–323
10. Bouzeghoub, M. and Peralta, V. (2004). A framework for analysis of data freshness. In IQIS'2004 (SIGMOD workshop), Paris, France

Type	International scope	National scope
Journals	8	2
Proceedings		1
Book chapters	1	
Theses	1 PhD thesis	1 Master thesis
Conferences	34	22
Posters and demos	4	3
Patents	1	
Keynotes and tutorials	3	1
Preprints	6 CoRR	4 technical reports

Scientific advising :

I co-supervised 6 defended PhD theses:

Mahfoud Djedaini	Automatic assessment of OLAP exploration quality (2017). French regional funding. Co-supervised with Patrick Marcel.
Krista Drushku	User Intent based Recommendation for Modern BI Systems (2019). Industrial funding (CIFRE, SAP company). Co-supervised with Nicolas Labroche and Patrick Marcel.
Frederick Bisone	Extraction de trajectoires sémantiques à partir de données multi-capteurs : application à des véhicules de secours (2021). Industrial funding (CIFRE, Petit Picot company). Co-supervised with Thomas Devogele and Laurent Etienne.
Clément Moreau	Fouille de séquences de mobilité sémantique (2021). French national funding (ANR project). Co-supervised with Thomas Devogele and Laurent Etienne.
Raymond Ondzigue Mbenga	Système d'information décisionnel, de la narration à la simulation : application à la surveillance épidémiologique de la tuberculose au Gabon (2023). International funding (World Bank). Cotutelle, co-supervised with Thomas Devogele and Edgar Ngoungou.
Faten El Outa	A framework for crafting data narratives (2023). French regional funding. Co-supervised with Patrick Marcel.

I am currently co-supervising 4 PhD theses:

Flavia Serra	Started in 2020 (defense planed at spring 2024). Uruguayan national funding. Cotutelle, co-supervised with Adriana Marotta and Patrick Marcel.
Raphaël Bres	Started in 2021. French regional and national funding (IGN). Co-supervised with Ana María Olteanu Raimond, Cyril de Runz and Arnaud Le Guilcher.
Hiba Merakchi	Started in 2023. French regional funding. Co-supervised with Thomas Devogele.
Guillaume Tejedor	Started in 2023. French local funding (Les Rabelaisiennes). Co-supervised with Hélène Blasco, Patrick Marcel and Nicolas Labroche.

I co-supervised the postdoctoral project of Kamal Boulil (2014-2015) and the postgraduate project of Alexandre Chanson (2020). I am currently co-supervising the postdoctoral project of Louise Parkin (2023-2024).

I supervised the master theses and internships of Imen Haddar (2023), Clément Legroux (2021), Willeme Verdeaux (2018), Oliver Ripka (2009) and Elena Martirena (2008), and co-supervised those of Antoine Chédin (2019), Abboubakar Sidikhy Diakhaby (2019), Yann Raymont (2018), Federico Mosquera (2015), Tahirou Famata (2012), Laura González (2009), Edmond Herri (2007) and Salvador Tercia (2006). I also supervised 12 engineering projects (one-year projects) between 2000 and 2009.

Participation to scientific defenses:

- Reviewer of the PhD thesis of Matteo Francia (University of Bologna, Italy, 2021).
- Examiner of the PhD defense of Redha Benhissen (University of Lyon 2, France, 2023).
- Examiner of the PhD defenses of the students I co-supervised: Faten El Outa (2023), Raymond Ondzigue Mbenga (2023), Clément Moreau (2021), Frederick Bisone (2021), Krista Drushku (2019) and Mahfoud Djedaini (2017).
- Examiner of mid-term PhD defenses of several local students (2018-2023) and international students of the IT4BI-DC program (2015-2017).
- Reviewer of the Master theses of Flavia Serra (2016), María Viola (2014) and Ignacio Larrañaga (2007).
- Examiner of the Master defenses of several international students of the IT4BI (2018) and BDMA (2019) Erasmus+ Master programs.
- Mentor at the Doctoral Consortium of the ADBIS conference (2023).

Participation to scientific projects:

2024-2027	IntForOut – Multisource spatial data INTegration FOR the monitoring of ecosystems under the pressure of OUTdoor recreation. French national funding (ANR).
2023	Educ'action – Ehics of Artificial Intelligence in Education and Training Sciences. French national funding (MaDICS, CNRS).
2023-2026	Data quality within data preparation for Big Data analysis. Uruguayan national funding (CSIC).
2022-2027	JUNON – Digital twins for natural resources. French regional funding (ARD).
2022-2025	Optimedias - Optimization of Data Exploration by Artificial Intelligence in Health. French regional funding.
2018-2022	Madona – Mastering Interactive Data Analysis for Journalistic narration. French national funding (MaDICS, CNRS).
2017-2021	Mobi'Kids – The role of urban educative cultures in the evolution of children's daily mobility and life context; collection and analysis of geolocated and semantically enriched tracks. French national funding (ANR).
2014-2017	DOPAn – Open data for monitoring and analysis. French regional funding.
2011-2014	Personae – New approaches to the representation of people, social networks and space (Middle Ages and Renaissance). French regional funding.
2009-2012	Codex – Efficiency, dynamism and composition for XML models, algorithms and systems. French national funding (ANR).
2009-2012	A platform for the construction and verification of composed modules. French local funding (Orleans-Tours collaboration). I was co-manager.
2008-2010	Data quality management for model improvement in genome-wide association studies. International funding (Microsoft Research).
2008-2010	Evolution and quality management in dynamic data integration systems. International funding (Stic-Amsud).
2006-2009	Quadris – Quality of data and multi-source information systems. French national funding (ANR).
2005-2007	Analysis of quality factors in multi-source information systems. Uruguayan national funding (CSIC).
2004-2007	APMD – Personalized access to data masses. French national funding (ANR).
2002-2004	Data warehouse logical design: techniques and tools. Uruguayan national funding (CSIC).
2002-2004	Metadata-based environment for the development of decisional systems: application to bio-maritime domain. Uruguayan national funding (DINACYT).

Participation to scientific committees:

- Program chair of French conference EDA (2021).
- Guest Editor of special issues of international journals ISF (2024), IJDWM (2014).
- Reviewer of several international journals: ECIS (2024), DKE (2023, 2021, 2020), IS (2022, 2017-2019), COMSIS (2022), JDIQ (2021, 2016), IJBIS (2016), IJDWM (2015), OJDB (2014), ISF (2011); and national journals: ISI (2014-2018).
- Member of program committee of several international conferences: DAWAK (2023, 2020), ADBIS (2023), DOLAP (2017-2020), ICEUTE (2013, 2020), ICIQ (2016), QMMQ (2014), QDC (2011), DEXA (2004-2008) ; and national conferences : BDA (2023), EGC (2023-2024), EDA (2022, 2013-2015), INFORSID (2017), QLOD (2016), CLEI (2009-2014), RJCRI (2006).
- Member of PhD award commettee of French conference BDA (2022).
- Reviewer of research projets for the Uruguayan Research Agency - ANII (2014, 2018).

Organization of scientific events:

- Co-chair of international workshop QAUCA (2019-2020) and French conference EDA (2021).
- Co-chair of a special session of the international conference ICEUTE (2020).

- Co-chair of French challenges “Predicting changes in groundwater level”, within the French conference EGC (2022) and “Narrating Rennes by the data”, within Madona workshop-MaDICS (2021).
- Co-chair of French regional workshops on “Data Sciences for Societal Challenges” (2023) and “1st Decision-Making Day in Center Region” (2015).
- Member of organization committee of international conference CIAA (2011) and French conferences BDA (2024), EGC (2022) and EDA (2013).
- Co-chair of the “Book donation” program of the international conference SIGMOD (2004).

Invitations and talks:

- Invited talk “Data makes the story - From Business Intelligence to Data Storytelling”, Workshop on Data Science for Societal Challenges, Blois, France, 2023.
- Seminar “Data Quality – Applications to the Health domain”, DEBIM/UREMCSE team, Libreville, Gabon, 2023.
- Invited talk “From source data to data narratives: accompanying users in the way to interactive data analysis”, Workshop on Human in the loop for data mining and machine learning (GdR DIAMS), Orleans, France, 2023.
- Talk “From source data to data narratives: accompanying users in the way to interactive data analysis”, Workshop on Computer Science at Center Region (JIRC), Orléans, France, 2022.
- Invited talk “From source data to data narratives: accompanying users in the way to interactive data analysis”, Workshop of the DOING action (GDR MaDICS), online, 2021.
- Keynote “From source data to data narratives: accompanying users in the way to interactive data analysis”, ADBIS/TPDL/EDA joint conferences, online, Augst 2020.
- Several invitations during mon sabbatical semester (CRCT) in 2020 (the 2 latter were postponed and replaced by virtual meetings, due to covid pandemic):
  - Polytechnical University of Catalonia, Barcelona, Spain, February 2020.
  - University of Ioannina, Ioannina, Greece, May 2020.
  - University of the Republic, Montevideo, Uruguay, July 2020.
- Invited talk “From data quality to analysis quality”, Workshop on Corpus for quality analysis of data exploration journalism (CAJOLE, CORIA-TALn conference), Rennes, France, 2018.
- Seminar “Assessing the Quality of Interactive Database Exploration”, ERIC team, University of Lyon 2, Lyon, France, 2018.
- Seminar “From data quality to data exploration quality”, University of Rennes 1, Rennes, 2016.
- Seminar (with Victor Baena Reina) “Similarity measures for career comparison”, Personae project, Tours, France, 2013.
- Seminar “Quality assessment platform Application to biomedical data”, Warehousing and mining seminar, Blois, France, 2009.
- Invited talk “Data Quality Evaluation in Data Integration Systems”, 2nd Workshop on Foundation on Databases and the Web (in honor to Alberto Mendelzon), Punta del Este, Uruguay, 2007.
- Seminar about my research works, Polytechnical University of Catalonia, Barcelona, Spain, 2004.
- Talk “A Framework for Assessing Data Quality in a Data Integration System”, 9th Workshop on Computer Science and Operational Research (JIIO), Montevideo, Uruguay, 2004.
- Talk (with Gonzalo Echagüe) “Image Management: Research and Implementation of a Solution”, 9th International Meeting of GeneXus Users, Montevideo, Uruguay, 2000.
- Talk “On the translation from the conceptual to logical schemes of Data Warehouses”, 6th Workshop on Computer Science and Operational Research (JIIO), Montevideo, Uruguay, 2000.
- Invited talk “Data Management Systems: Architectures and Integration of Heterogeneous Information in Data Warehouses”, “EI-business’2000” Workshop, Montevideo, Uruguay, 2000.

5 Other activities and responsibilities

Administrative responsibilities:

- Head of Computer Science department, Faculty of Sciences (from 2021).
- Member of the bureau of the Faculty of Sciences, in charge of business relationships (from 2017). I am also in charge of business relationships in Computer Science department (from 2010).
- Member of the Scientific Board of the ICVL Federation (from 2022).
- Elected member of the Computer Science Scientific Committee - CSDP (from 2021).
- Responsible of Computer Science Bachelor program for next contract (2024-2028). I coordinated the development of the program and ensured its sustainability. I also participated in the development of Bachelor programs for previous contracts (2013-2017 and 2018-2022), as well as for Master programs DS4SC (2024-2028), BDMA (2018-2022), SIAD (2013-2017) and IT4BI (2013-2017).
- Responsible of a school year of a Bachelor or Master program in Computer Science (2009-2021). I took care of L1 and L2 (2009-2010), L3 (2010-2013), M2 SIAD (2013-2017), M1 BDMA (2017-2020) and M2 BDMA (2020-2021).
- Co-responsible Erasmus, in charge of Computer Science mobility (2019-2021).
- Local coordinator of the European Computer Science – ECS – Erasmus program (2019-2021). This program coordinates mobility of Bachelor students within a network of partner universities. I participated in the annual meetings of the steering committee (2009-2021) and in the set-up of the EMACS Master which followed the Bachelor program (2009-2011).
- Elected member of the council of LIFAT laboratory (2018-2019).
- Member of the bureau of BdTln research team, in charge of the web site of the team (2015-2019).
- Member of selection committees for the recruitment of Lecturers in Blois (3 committees 2014-2024), Paris (2023), Lyon (2021), Bourges (5 committees 2015-2018) and Orléans (2015).
- Member of the steering committee of 2 international Erasmus+ Master programs (2011-2019): IT4BI and BDMA. I participated in set-up projects, annual meetings of the steering committee, candidate selection meetings and Master theses defenses.
- Member of internship exam board of many Bachelor and Master students (from 2008). In addition, I am responsible for coordinating and monitoring interns as well as organizing defenses (from 2010). I regularly organize conferences and seminars where companies present a sector of application, a technology or describe their jobs to Bachelor and Master students. I also coordinate CV, interview and job dating workshops (led by HR managers from local companies or by MOIP project managers).
- I also took part in collective activities during my thesis and postdoc at the University of Versailles (2003-2007) and at the start of my career at the University of the Republic, Uruguay (1996-2008). In particular, I participated in the set-up and administration of websites, the installation and administration of computer networks and the organization of meetings, seminars and workshops.

Teaching activities

My teaching concerns Bachelor and Master courses, in French, English and Spanish, mainly around databases, data warehouses, data quality and project advising.

Here is a summary of my courses throughout my career:

University of Tours, France (since 2008)	Databases, data warehousing, data quality, personalization, business intelligence, personalization, semantic web, information retrieval, programming, computer networks
University of Paris Descartes, France (2015-2021)	Data quality
University of the Republic, Uruguay (1996-2008)	Databases, information systems, data warehousing, data quality, software engineering, programming, computer networks
University of Buenos Aires, Argentina (2008)	Data warehousing
University of Versailles, France (2003-2007)	Advanced databases, decision-support systems, data integration, software engineering
South Autonomous University, Uruguay (1996-2003)	Information systems, data warehousing, decision-support systems, programming, computer networks, computer architectures
8th Computer Science Summer School, Río Cuarto, Argentine (2001)	Data warehousing