A framework for crafting data narratives

Faten El Outa

UNIVERSITÉ DE TOURS

ÉCOLE DOCTORALE : MIPTIS
***ÉQUIPE*: BDTLN**

THÈSE présentée par :

Faten EL OUTA

soutenue le : **21 Décembre 2023**

pour obtenir le grade de : **Docteure de l'Université de Tours**
Discipline / Spécialité : **Informatique**

A framework for crafting data narratives

THÈSE dirigée par :

Mr. MARCEL Patrick	Professeur, Université d'Orléans

RAPPORTEURS :

Mme. SI-SAID CHERFI Samira	Professeure, CNAM de Paris
Mr. MATE Alejandro	Professeur, Université d'Alicante

JURY : ***(directeurs, rapporteurs, examinateurs)***

Mr. MARCEL Patrick	Professeur, Directeur, Université d'Orléans
Mme. PERALTA Veronika	Maître de conférence, Encadrante, Université de Tours
Mme. SI-SAID CHERFI Samira	Professeure, Rapportrice, CNAM de Paris
M. MATE Alejandro	Professeur, Rapporteur, Université d'Alicante
Pr. VASSILIADIS Panos	Professeur, Examinateur, Université de Ioannina
Mr. VENTURINI Gilles	Professeur, Président du Jury, Université de Tours
Dr. CHAGNOUX Marie	Maître de conférence, Examinateur, Université de Paris 8

Dedication

To my beloved family and friends,

you have always been there for me with love, support, and fun. Your being around has made my days brighter and filled with hope.

—

To my extended family, friends, supervisors and teachers,

your help and belief in me made me more confident and motivated. Your advice, support, and feedback helped me in my studies and personal development.

—

For everyone who has brought me a smile or a kind word through hard and soft times,

may this heartfelt dedication be remembered with respect to the deep and profound effect you have on the lives of others.

—

Acknowledgements

I am very thankful to God for all the good things that have happened during my PhD studies. I'm grateful for my good health and happy moments that made the journey more enjoyable.

I would like to express my deepest appreciation to my supervisor and advisor for their unwavering guidance, encouragement, and mentorship throughout my doctoral studies. Their expertise and insights have been invaluable, shaping not only my research but also my personal growth.

I am indebted to the members of my committee for their constructive feedback and dedicated time invested in reviewing my work.

To my colleagues at the LIFAt laboratory, particularly the BDTLN group, I extend my heartfelt gratitude for their collaboration, encouragement, and camaraderie. Your support has been instrumental in overcoming challenges.

I am profoundly thankful to my family for their unconditional love, unwavering belief in my abilities, and constant encouragement. Their support has been my source of strength throughout this journey.

To my friends, both near and far, I am grateful for your unwavering support, understanding, and encouragement. Your friendship has been a source of joy and inspiration.

Lastly, I express my deepest gratitude to my partner, Hussein, whose unwavering support, love, and companionship have been my anchor throughout this journey.

I want to thank God once more because I've achieved everything I aimed for: finishing my PhD thesis, getting a job, and finding love.

Publications

No.	Cited by	Year	Publications
1	**26**	2020	F El Outa, M Francia, P Marcel, V Peralta, P Vassiliadis. *Towards a conceptual model for data narratives.* ER
2	**9**	2020	F El Outa, M Francia, P Marcel, V Peralta, P Vassiliadis. *Supporting the generation of data narratives.* ER Forum/Posters/Demos
3	**5**	2022	F EL Outa, P Marcel, V Peralta, R Da Silva, M Chagnoux, P Vassiliadis. *Data narrative crafting via a comprehensive and well-founded process.* ADBIS
4	2	2021	RO Mbenga, V Peralta, T Devogele, F El Outa, SM Nzondo, EB Ngoungou. *Processus de narration de données en intelligence épidémique avec application à la pandémie de tuberculose au Gabon.* JCIM
5	1	2022	RO Mbenga, V Peralta, T Devogele, F EL Outa, S Maghendji, N Edgard. *A data narrative about tuberculosis pandemic in Gabon.* DARLI-AP
6		2022	A Chanson, F El Outa, N Labroche, P Marcel, V Peralta, W Verdeaux. *Generating Personalized Data Narrations from EDA Notebooks.* DOLAP
7		2021	M Chagnoux, R da Silva, F El Outa, N Labroche, P Marcel, V Peralta. *Modéliser la démarche du data journaliste.* H2PTM
8	1	2021	F El Outa, P Marcel, V Peralta. *Un modèle conceptuel de narration de données.* EDA
9		2023	P Marcel, V Peralta, F El Outa. *A declarative approach to data narration.* CoRR
10		2023	F El Outa, P Marcel, V Peralta, P Vassiliadis. *Highlighting the Importance of Intentional Aspects in Data Narrative Crafting Processes.* Information Systems Frontiers
11		2024	P Vassiliadis, P Marcel, F El Outa, V Peralta, D Gkitsakis. *A Conceptual Model for Data Storytelling Highlights in Business Intelligence Environments* Arxiv

Résumé

Le monde est rempli d'une quantité écrasante d'informations, et en faire sens constitue un défi significatif. Les récits ont depuis longtemps été reconnus comme des outils puissants de communication. Cette thèse de doctorat explore le domaine des récits soutenus par des données, en introduisant un nouveau cadre conçu pour élaborer des récits de données.

La narration de données est généralement décrite comme l'activité d'élaborer des récits étayés par des faits extraits de l'exploration et de l'analyse des données, en utilisant des visualisations interactives. Malgré l'intérêt croissant pour la narration de données dans plusieurs domaines (par exemple, le journalisme, les affaires, l'e-gouvernement), il n'existe pas de définition consensuelle de la narration de données, sans parler d'un modèle conceptuel ou logique. De plus, le processus d'élaboration de récits de données est peu documenté et n'a pas encore été formellement décrit.

Cette recherche vise à relever les défis liés à la transition des récits traditionnels vers les récits de données en proposant un cadre pour élaborer des récits de données. Le cadre proposé comprend deux composantes clés : (i) un modèle conceptuel qui guide les recherches ultérieures et facilite la compréhension, la normalisation, la réutilisation et le partage des récits de données ; (ii) un processus illustrant les différentes phases et activités impliquées dans la narration de données, notamment l'analyse des données, l'extraction des messages pertinents, la structuration des messages en une histoire cohérente et la représentation visuelle ; (iii) le cadre affine davantage le modèle conceptuel pour les récits de données, en se concentrant spécifiquement sur l'exploration des cubes de données en tant que cas d'utilisation distinct dans le domaine de la narration de données. Ce modèle affiné a le potentiel d'aider les narrateurs de données à naviguer dans des ensembles de données multidimensionnels complexes, à extraire des informations précieuses et à prendre des décisions éclairées.

De nombreuses expériences menées soulignent l'applicabilité pratique du modèle conceptuel proposé pour les récits de données et mettent en avant l'importance d'utiliser un processus structuré dans l'élaboration de récits de données. Ces expériences illustrent également comment le processus proposé intègre des activités issues de recherches antérieures.

En proposant ce cadre, cette recherche vise à contribuer au développement d'une approche unifiée de l'élaboration de récits de données, favorisant une meilleure compréhension, normalisation et collaboration dans le domaine. De plus, notre cadre crée des possibilités de recherche nouvelles et intéressantes, encourageant l'exploration de nouveaux territoires dans la narration de données. Ces modèles proposés constituent une première étape cruciale dans le développement du domaine. En fournissant une orientation et en automatisant des activités difficiles ou complexes dans le processus de narration de données, ils peuvent ouvrir la voie à la création de technologies et d'outils facilitant

l'élaboration de récits de données. Cette thèse de doctorat fait progresser la compréhension de la narration de données, fournissant des informations précieuses aux praticiens et aux chercheurs engagés dans la création et l'analyse de récits basés sur les données.

Mots-clés: Récit de données, narration de données, narration, modèle, processus, analyse, structure, visualisation, et intention.

Abstract

The world is filled with an overwhelming amount of information, and making sense of it all is a significant challenge. Narratives have long been recognized as powerful tools of communication. This Ph.D. thesis delves into the realm of narratives supported by data, introducing a novel framework designed for crafting data narratives.

Data narration is typically described as the activity of crafting narratives supported by facts extracted from data exploration and analysis, using interactive visualizations. In spite of the increasing interest in data narration in several communities (e.g. journalism, business, e-government), there is no consensual definition of data narrative, let alone a conceptual or logical model of it. Furthermore, the process of crafting data narratives is loosely documented and has not yet been formally described.

This research aims to tackle the challenges associated with the transition from traditional narratives to data narratives by proposing a framework for crafting data narratives. The proposed framework includes two key components: (i) a conceptual model that guides further research and facilitates understanding, standardization, reuse, and sharing of data narratives; (ii) a process illustrating the diverse phases and activities involved in data narration, including data analysis, extraction of relevant messages, structuring of messages into a coherent story, and visual representation; (iii) the framework further refines the conceptual model for data narratives, specifically focusing on data cube exploration as a distinct use case within the realm of data narration. This refined model holds the potential to aid data narrators in navigating intricate multidimensional datasets, extracting valuable insights, and making informed decisions.

Many conducted experiments underscore the practical applicability of the proposed conceptual model for data narratives and emphasize the significance of employing a structured process in crafting data narratives. These experiments also illustrate how the proposed process integrates activities from prior research.

By proposing this framework, this research aims to contribute to the development of a unified approach to crafting data narratives, leading to better understanding, standardization, and collaboration in the field. Furthermore, our framework creates interesting fresh possibilities for research and acts as an encouragement for exploring new territories in data narration. These proposed models serve as a crucial first step in the development of the field. By providing direction and automating difficult or complex activities in the data narrating process, they may open the way for the creation of technologies and tools that facilitate the crafting of data narratives. This PhD thesis advances the understanding of data narration, providing valuable insights for practitioners and researchers engaged in the creation and analysis of data-driven narratives.

Keywords: data narrative, data narration, data story, storytelling, model, process, analysis, structure, visualization, and intention.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Context. "*Stories are the creative conversion of life itself into a more powerful, clearer, more meaningful experience. They are the currency of human contact*", said Robert McKee. This quote emphasizes the importance of narratives because they make our lives better, help us learn more, and bring us closer together. They make us feel things, organize knowledge, and help us understand other people's points of view. Narratives have a big effect on how we grow as people, how we connect with others, and how we make sense of the world. *If narratives have the power to reshape the world, imagine the profound impact when narratives are supported by data.*

Importance of narratives. At their core, narratives, as a fundamental aspect of human culture, serve various purposes such as entertainment, education, cultural preservation, and moral instruction [16, 48]. Traditional narratives refer to the conventional narration that relies on language, characters, plot, and subjective interpretation to convey a message or tell a story. Traditional narratives often emphasize emotional and personal experiences, and they allow for creative imagination and subjective perspectives. But when data are incorporated into narratives, they become much more powerful. When statistics and findings back up a narrative, it makes it even more interesting and convincing. Data provides a solid base of proof, giving the narrative more credibility and making a powerful mix of feeling and logic that appeals to both the heart and the mind.

However, in the current era characterized by a tsunami of data, the effective communication of findings derived from data analysis has become increasingly critical. To address this need, data narratives are viewed as a way for incorporating finding retrieved from data into compelling and understandable narratives. Data narratives bridge the gap between data analysis and narrative by integrating elements of narrative with data visualization and analysis. This transition necessitates a fundamental shift in approach, as raw data is transformed into meaningful and engaging narratives.

Data narratives. Data narration, i.e., narrating with data visualizations [80], is considered as the activity of producing narratives supported by facts extracted from data analysis, using interactive visualizations [28]. More concretely, 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 [92]. Data narration exploits not only data analysis and statistics but also data visualization, qualitative and contextual analysis, and presentation [147]. Further, it involves selecting the most relevant data, determining the appropriate visualization techniques, and weaving together a cohesive narrative that engages and resonates with the audience [124]. Data

narrative is also an effective way to communicate data-driven insights to non-technical audiences. By presenting data in a narrative format, data narratives make information more accessible and relatable, fostering a deeper understanding and engagement [26].

Importance of data narratives. In the modern era, the emergence of data narrative, also known as data storytelling, has brought together different communities of practice and research in exploring the potential of conveying narratives through data [79]. As a result, disciplines such as journalism [77], industry [111], e-government [134], data management [1], data visualization [57]. and computer-human interaction [21] have increasingly recognized the importance of data narrative in their respective fields. Overall, data narrative enhances communication, understanding, and decision-making across various fields and disciplines.

Over the years, data narratives have garnered significant attention and research in this important field. In this domain, a plethora of terms spanning multiple fields have emerged, including visual narratives, data stories, data storytelling, etc. The key characteristics of data narratives have been the subject of numerous studies [19,125,182]. This collaborative effort highlights how crucial it is to comprehend and make use of the potential of data narratives in a world that is becoming more and more data-centric.

Data narration converts raw data into valuable findings, filling a critical gap in the fields of data science and reporting. It offers the narrative context necessary for comprehending complex analytical results in the field of data science, assisting in the interpretation of data patterns and their implications. Data narratives, when it comes to reporting, fill the gap between technical findings and non-technical audiences, ensuring that information is not only effectively conveyed but also engages and informs decision-makers.

In the field of data narration, the exploration of data cubes represents a significant and compelling use case. The importance of data narration in data analysis is highlighted by the vivid illustration of the difficulties and opportunities involved in communicating insights from complex multidimensional datasets.

Data narratives are commonly used by various professionals, including journalists, scientists, and other communicators such as marketers, public health officials, as a means of conveying compelling messages to a specific audience.

Example. Any sort of narration that is constructed based on data might be considered a data narrative. A data narrative could take the form of a data video, a fact sheet, a news article, a notebook, an infographic or any other visual format aiming to communicate messages. An example of a data narrative is depicted in Figure 1.1[1]. This data narrative offers a clear and easily comprehensible presentation of pricing information in Lebanon. It encompasses a range of messages that pertain to various aspects of the cost of living, such as food prices, restaurant expenses, beer costs, accommodation rates, transportation expenses, and the overall expenditure for a one-week stay in Lebanon for different types of tourists. This example serves as a practical demonstration of how data narratives can effectively convey valuable information to a diverse audience.

Additional examples of data narratives are presented in section 1.1 of Annex A demonstrating the utilization of diverse data sources to communicate complex information through different forms of data narratives in an understandable and engaging manner.

[1] https://hikersbay.com/prices/lebanon?lang=en

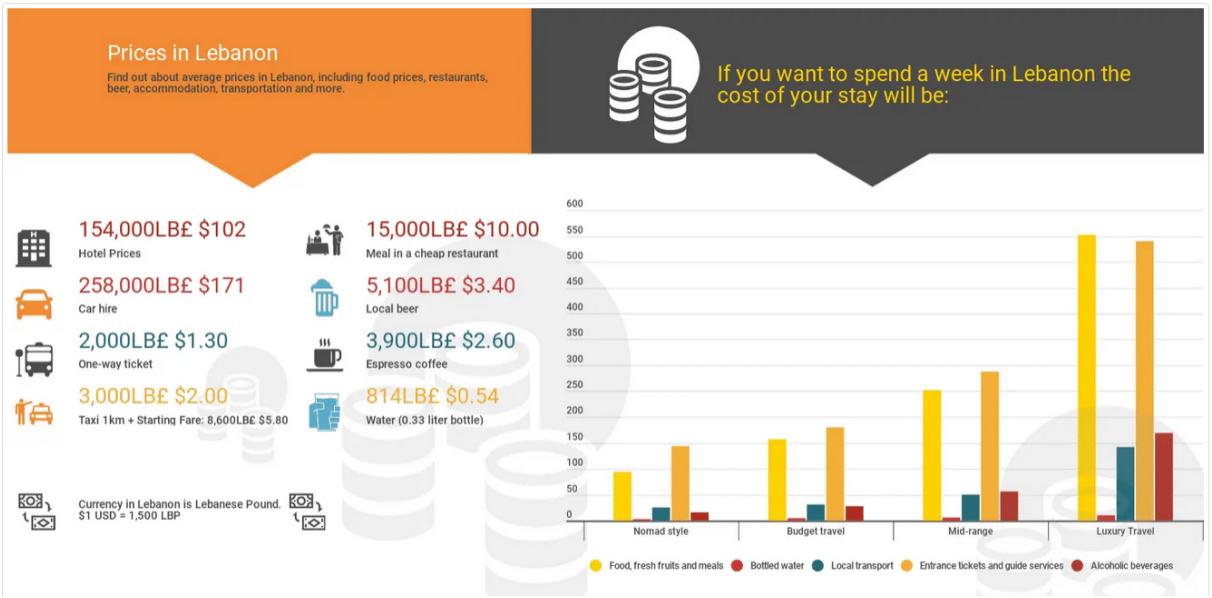


Figure 1.1: Prices in Lebanon, including food prices, restaurants, beer, accommodation, transportation and more.

Crafting such data narratives raises numerous research questions, particularly in regards to identifying the fundamental concepts and determining the initial steps of this journey. Data narrators may find themselves thinking about the necessary skills and the sequence of activities required to create a compelling data narrative. Data narrators may ask also if they should be skilled at data exploration, storytelling, creating visuals, etc.

Objective. In light of the diverse and multifaceted nature of the data narratives domain, it is crucial to propose a framework for crafting data narratives that includes the main concepts and activities coming from diverse domains, ensuring a holistic approach to data narratives and facilitating the data narrator's crafting. Achieving this objective requires to answer the following some questions:

- (RQ1) How can data narratives be defined and modeled, considering the novelty and heterogeneity of the data narratives domain?
- (RQ2) What are the key components of a framework for constructing data narratives?
- (RQ3) What are the primary activities involved in the process of crafting data narratives?
- (RQ4) How to account data narration for the particular and typical case of data cube exploration?

These research questions aim to address the challenges related to defining, modeling, and creating data narratives, while also emphasizing the importance of providing guidance and support to narrators. Additionally, the questions focus on the development of a framework that can enhance the construction of data narratives and facilitate communication of complex information. Such framework for crafting data narratives plays a critical role by providing essential guidance to data narrators, increasing the overall quality of the narratives, and increasing the effectiveness of conveying findings from data analysis.

Contributions The PhD thesis makes significant contributions to the field of data narratives by addressing the aforementioned research questions, ultimately leading to the development of a comprehensive framework. The contributions can be summarized as follows:

- *Development of a Conceptual Model*: This thesis introduces a conceptual model for data narratives, which serves as a foundational model for understanding the diverse aspects and core concepts within the field. This model aids in standardization, reuse, and sharing of data narratives, providing a structured approach for practitioners and researchers.

- *Development of an Activity Model*: This research presents a well-defined process for crafting data narratives, encompassing multiple phases. A flow diagram illustrating the diverse activities involved in data narration, including data analysis, extraction of relevant messages, structuring of messages into a coherent narrative, and visual representation.

- *Illustrating a typical use case of data narrative in the context of data cube exploration*: This thesis refines the data narrative conceptual model accounting for data cube exploration as a specific case of data narration. By considering various elements such as data cube environment, belief of data narrator and the history of data narrative crafting, the research models a part of data narrative within the context of data cube exploration.

Overall, these contributions address the lack of a holistic framework for data narratives, considering the heterogeneity and novelty of the domain.

Thesis structure overview. Figure 1.2 proposed an order for reading the manuscript while grouping research questions into their respective chapters to address them effectively. The blue arrows represents a sequence between chapters and black arrows represents chapters that might go to the Appendix for more details. The thesis is structured as follows:

Chapter 2 provides an exploration of the current state of the art in data narrative by examining its related domains. It offers key takeaways to facilitate understanding of the domain. Chapter 3 answers (RQ1, RQ2) and focuses on presenting the conceptual model for data narratives, outlining its fundamental concepts. From this chapter, the reader can proceed to the Appendix A to acquire further information about the practical implementation of data narrative concepts, as documented in the literature. Chapter 4 answers (RQ3) and delves into the process of crafting data narratives, providing detailed activities and best practices for creating compelling narratives. Chapter 5 answers (RQ4). It is dedicated to refine a part of data narratives model in the context of data cube exploration. Readers can move on to Appendix A from this chapter to access additional examples. Chapter 6 concludes the thesis by summarizing the main findings and contributions and suggests future directions to further enhance the framework for crafting data narratives.

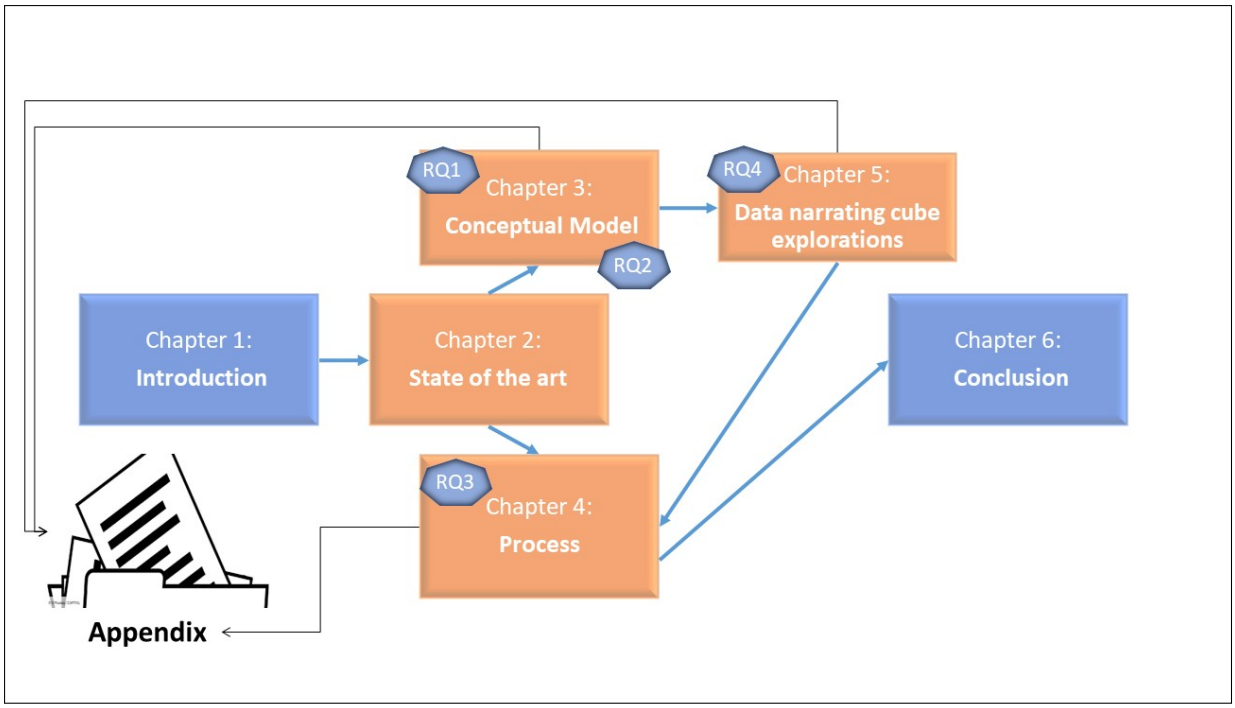


Figure 1.2: Plan for my PhD thesis

Notices for reading In what follows, we use the terms *author*, *analyst* or *data narrator* for the designer of the data narrative. The author is not necessarily a business analyst, they can be a data journalist, or a plain data enthusiast, aiming to produce a report of findings. We also assume an *audience*, *reader* for the produced outcome, which includes the people that will see or read the data narrative.

Chapter 2

State of the art

Data narration refers to the practice of telling stories using data. Data narration involves understanding the data, turning data into messages and communicating these messages in the most meaningful and compelling way. Such tasks benefit from a range of techniques drawn from diverse domains like *data exploration* and *analysis*, which involves examining the data to extract relevant messages, *data visualization*, which entails creating visual representations to effectively communicate the crafted messages, *narrative*, which involves identifying the most impactful way to convey the messages, and more.

In this chapter, we will examine the various definitions of data narratives in order to obtain a better understanding of the subject. We also identify the concepts and tasks that contribute to the development of a data narrative. Through the content presented, readers can learn key points for understanding what is and how to craft a data narrative. Specifically, this chapter answers questions about What is a data narrative? What are the components of data narrative? How does one create a data narrative?

Contents

2.1 From narratives to data narratives

While using many terms (e.g., narrative visualization, visual storytelling, data driven storytelling), the data visualization community has recently brought much attention to data narration [28]. It is important to address these terms such as data storytelling and visual storytelling and provide concise definitions, while recognizing that there is no formal definition or conceptual framework in these areas.

Within the domain of narrative modeling, various theoretical frameworks have been proposed, including the ones mentioned earlier [53] but none of them qualifies for data narratives. This is primarily due to the absence of content preparation derived from data. While these frameworks may offer valuable insights into narrative structure, they do not incorporate the essential element of data analysis and its result to shape the narrative content. Even though the notion of data is not present in these models, some aspects of classical narration theory, as described e.g., by Chatman [32], should be studied to understand the fundamental structure of narration and draw inspiration for the modeling of data narratives. This is the topic of Subsection 2.1.1. In Subsection 2.1.2, several definitions of data narratives are examined with a particular focus on the disparity between narratives and data narratives.

2.1.1 Narratives

Narrative models are theoretical frameworks used to represent and analyze the structure, content, and dynamics of narratives. These models often aim to capture the key elements, relationships, and patterns within a narrative to facilitate a deeper understanding of its storytelling aspects. One common approach to narrative modeling is the two-level framework proposed by Todorov et al. [160], which distinguishes between the plane of story content called *histoire* (story), and the plane of style and point-of-view referred to as *discourse*. Another perspective, introduced by Bal [11], involves a three-level distinction. Bal's model includes the *fabula*, which they define as a series of logically and chronologically related events that are caused or experienced by actors; the *story*, in which a narrator (perceiver) selects some elements of the fabula to convey and omits others; and the *text*, where words are chosen to convey the story in a discourse.

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). In the most widely used structuralist terminology, the answer to the "what" question is called a **story** and the answer to the "how" question is called a **discourse** [4].

Chatman [32] distinguishes narration's elements based on the what and how questions, defining narrative as a couple of story (content of the narrative) and discourse (expression of it). Both, story and discourse, have a form and a substance (see the four sub-levels in Figure 2.1). In more details, the story is the fundamental element of the narrative, consisting of the events, characters, and setting that make up the content of the story. It encompasses the overall meaning of the narrative and provides the foundation for the other elements of the narrative structure. The discourse is the actual text or language used to convey the story to the reader or audience. It includes elements such as dialogue, description, and narration, and is shaped by the author's stylistic choices and the

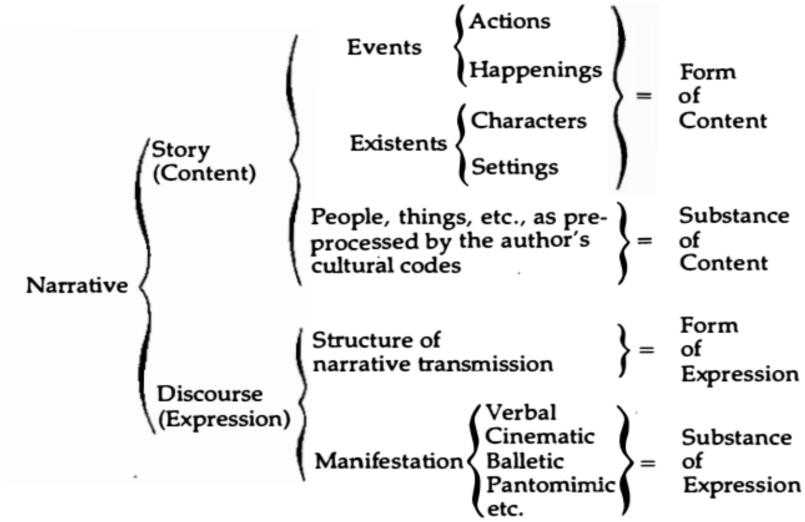


Figure 2.1: Narrative structure from [32]

conventions of the genre.

The story has a form, that is the set of possible objects, events, etc., and a substance, which is a **composition** of story elements (i.e., events, settings, behaviors, characters) as pre-processed by the author's cultural code. A discourse has a form, which is a translation of the story content to a **structured combination** of the story elements. In other words, this means that, out of the entire story as it actually happened, when constructing a discourse, the author picks an interesting subset to present. The discourse also has a substance that includes the set of all **media** used to show structured elements. The substance of discourse is the material nature of the linguistic elements, for example, the actual sounds made by voices, or marks on paper.

Take away. The narrative model of Chatman [32] emphasizes the distinction between story and discourse and ultimately describes four different layers. In summary, the story can be seen as content of the narrative, while the discourse is its presentable manifestation, obtained through author's editions. These four layers are as follows: (i) the story form provides the foundation for the narrative's content, (ii) the substance of the story involves the composition of story elements, which have been pre-processed by the author's cultural code, (iii) the discourse form is the structured representation of the story content in the narrative, and (iv) the substance of discourse includes the media and material nature of the linguistic elements used to convey the narrative.

2.1.2 Data narratives

In this section, we intend to build an extensive list of data narratives-related definitions, covering all of its various terms. These terms span across various domains, including but not limited to those such as data story, visual narrative, data storytelling, etc. By engaging in this study, we aim to understand what is a data narrative and identify the essential *tasks* that contribute to the development of a data narrative. It's important to emphasize that a "data narrative" serves as the end result, while "data narration"

envelops the process of developing and delivering that narrative. In practice, these terms are intimately connected and are frequently interchanged.

Informal definitions In contrast to traditional narration, data narration is the fact of narrating with data [181] using different visual means [3, 80, 122]. Carpendale et al. [28] defined data narration as the activity of producing narratives supported by facts extracted from data analysis, using interactive visualizations. Kosara and McKinley [92] intuitively define a data narration as an ordered sequence of steps, each of which primarily consists of visualization, which can include text and images but essentially are based on data. The authors note that journalists work with a model of story construction where the order of events is consistent and clear, for the story to be comprehensible. Shi et al. [150] proposed a logical definition of data narrative. They defined data narrative as a sequence of meaningfully connected data facts [156, 168], a numerical property that pertains to a particular data subspace, that are ordered according to narrative logic. In this context, "narrative logic" refers to the organization and ordering of these data facts in a manner that creates a coherent and meaningful data narrative. For example, Datashot's [168] narrative logic includes topical segmentation, evaluating topics by relevance, selecting the most relevant facts for each topic, and applying a similarity-based filter to avoid very similar facts in the final narrative.

Segel and Heer [147] refer to data narratives as narrative visualization, i.e., visualizations intended to convey stories, and emphasized the difference between traditional narrative and data narrative, namely the interactive potential of the latter. The authors insist that the notion of a chain of causally related events is central to the definition.

Process Crafting a data narrative encompasses several essential activities that revolve around three phases concerning data analysis, data visualization, and narrative structuring [46]. These activities are interconnected and contribute to the creation of a compelling data narrative. For example, in [92], journalists **collect information, which gives them the key facts**, and then they tie those facts together into a story. The authors note that the **goals**, tasks and tools used during the research phase differ from those in the writing phase, and that **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.

Before Chen et al. [35], data narration has been associated with activities encompassing both data analysis and storytelling while storytelling refers to data visualization techniques to visually communicate information to the audience. Data analysis phase requires to **see all aspects of complex data**, **explore their interrelationships**, and is supported by multiple coordinated views and sophisticated interaction techniques, while, data visualization phase is meant to convey only interesting and/or **important finding** extracted through the analysis, presented in a simple and easily understandable way. Data analysis and storytelling phases differ in their purposes, target users, kind of information dealt with, and methods of presenting the information and interacting with it. To support data narrative structuring, Chen et al. [35] proposed an intermediate phase between analysis and storytelling, in which the analyst **assembles and organizes information pieces** to be communicated.

Mosconi et al. [121] emphasized that data narration encompasses activities related to

the three aforementioned phases. These activities include: (a) explaining the context, which is determined by the author of the narrative who is responsible for identifying the relevant contextual factors; (b) identifying a coherent narrative to share values and norms, promote commitment and trust, communicate implicit knowledge, and establish emotional connections; and (c) providing effective visualization.

Lee et al. [98] also highlighted the involvement of the three phases in data narration activities, but they proposed a different order for these activities compared to previous works. In more details, they present data narration in the following manner: (a) a data narrative comprises a collection of **story pieces that are specific findings supported by data**; (b) the majority of these **story pieces are depicted visually** to reinforce one or more desired **messages**. The visualizations involve annotations such as labels, pointers, and text, or a narration to clearly **emphasize and accentuate the message** and avoid any ambiguity; and (c) the **story pieces are arranged in a meaningful order** or have a connection between them to **promote the author's overall communication objective**. This objective may vary from providing an illustration of facts to educate or entertain viewers, to offering thought-provoking opinions to convince or persuade them.

Take away. Data narration refers to the practice of visually communicating messages supported by data, distinguishing it from traditional narrative approaches. Tasks that revolve around the data analysis, structuring, and data visualization are the essence of data narration [28, 35, 46, 80, 121]. Crafting a data narrative involves a series of activities with the goal of extracting meaningful findings from data. These activities encompass the analysis of data to uncover key findings, the formulation of messages supported by findings, the arrangement of messages in a in a certain order, and the visual communication of this organized messages. In crafting a data narrative, the data narrator's intent becomes noticeable through the setting of a goal and the formulation and selection of pertinent messages to be conveyed to the intended audience.

2.2 Concepts of data narratives

Data narration involves the tasks of analyzing and exploring data in order to extract meaningful findings and communicating messages supported by findings in a structured and visually engaging manner, all while reflecting the intentions of the data narrator.

We observed that these tasks align with the four levels outlined in Chatman's [32] narrative model, even though the notion of data is not explicitly incorporated within those levels. To illustrate, the tasks align with Chatman's levels as follows: (i) *form of content* concerns the data analysis and exploration to retrieve findings among data, (ii) *substance of content* concerns the intention of data narrator in terms of formulating goals and messages, (iii) *form of expression* concerns the structuring and ordering of messages in a compelling way, and (iv) *substance of expression* concerns the visualization and the communication of messages to facilitate the audience's reception.

Based on these tasks, various concepts come to light, offering valuable insights to model the domain of data narratives. In this section, we will organize the state-of-the-art of data narrative's concepts based on these levels.

2.2.1 Form of content: Data exploration

The fact of exploring and analyzing data, as well as retrieving discoveries among data, plays a vital role in preparing the form of a story. It enables the data narrator to identify and select the most fitting story elements. One typical difference between narrative and data narrative is that the story elements are supported by data in the latter. Subsequently, we will delve into a more detailed examination of approaches that facilitate the generation of significant discoveries.

Exploration. Data exploration acts as the vital point for retrieving findings and patterns in the data. Data is a resource that is dormant and waiting to be explored for its full potential. Data exploration is notoriously described as a tedious activity because exploration tasks require profound analytical skills, experience, and domain knowledge.

Various approaches [67,72,83] exist for *supporting interactive data exploration*, ranging from the adaptation of the database layer to enhance its suitability for interactive data exploration (adaptive indexing, adaptive loading, adaptive storage, flexible architectures, architectures tailored for approximate processing) to the use of specific middleware to improve data exploration (data prefetching, query approximation, etc.) to the improvement of the user interaction to enhance result visualization and create more user-friendly exploration interfaces (result visualization, exploration interface). In addition, benchmarks are emerging [50, 137] to evaluate the performance of database systems for interactive data exploration workloads. These benchmarks provide a standardized way to assess the capabilities of database systems in handling the demands of interactive data exploration.

Collector. When it comes to data exploration, a data request is the key player in charge of collecting data from various sources. Data request has evolved significantly, transitioning from traditional analyses that demanded extensive expertise in SQL and programming to the modern platforms (like Tableau or Knime). Modern platforms facilitate composite analysis processes, interweaving actions of multiple types (e.g., SQL-like operators, OLAP multidimensional aggregations, visualization, model construction).

Nevertheless, as a starting point, authors of [7] identify many low-level actions of analytical activity. These low-level actions are typically the first steps in the data analysis process and serve as the foundation upon which more advanced and complex analytical techniques and methods can be applied. According to [118], analytical actions can be classified in three main categories: data retrieval actions, performed to select and filter the relevant data objects for the current assignment, data representation operations, performed to alter the point-of-view of the data objects (this includes OLAP cube operations), and data mining tasks, such as clustering and outliers detection.

Finding. One of the key results of the data exploration is the findings, which illuminates the inherent value and patterns found during the exploration journey led by data collectors. Characterizing meaningful findings in data has attracted a lot of attention, from the seminal work on discovery driven exploration [143, 144] and on knowledge discovery in databases [2]. Often, this characterization takes the form of *interestingness* scores for retrieved data [107] or *patterns* [17, 64].

Chen et al. [35] stress the importance of the relationships between insights for the selection of the more important ones. Discovery can encompass various relationships, such as values, differences, proportions, trends, categorizations, distributions, rankings, associations, extremes, and outliers [150].

Authors also highlight the significance of *insights*, where significance is defined in terms of statistical tests [67, 75, 84, 157, 179] or information theory [150]. Furthermore, recent proposals [149, 150, 168] deal with the design, structuring and formalization of discoveries in order to estimate their importance.

2.2.2 Substance of content: Data narrator's intention

In narratives, the composition of story elements, such as events, settings, behaviors, and characters, is shaped by the author's cultural code [32]. Similarly, in data narratives crafting, the data narrator's intention is evident through various elements. The data narrator establishes a clear goal, reflecting the intention of the data narrator. Analytical questions are formulated to guide the exploration of the data. These questions assist with determining the focus of the analysis. Additionally, the data narrator crafts and selects pertinent messages that effectively convey the findings to the intended audience and simultaneously answers an analytical question.

Within the context of data exploration and OLAP environments, data narrator intention modeling is set into motion. The construction of a user model that incorporates beliefs, objectives, and past interactions becomes essential for a comprehensive evaluation of the interestingness of cube queries within the multidimensional and contextual landscape of an OLAP environment [70]. In addition, a framework [107] is proposed for assessing the level of interest, as indicated by the degree of attention a particular piece of information receives depending on the data narrator's intention.

We organized the works based on the substance of content according to goals, analytical questions and messages, which reflect the data narrator's intent.

Goal. Since it provides direction for data exploration, the main goal is clearly present in data narratives.

Both the visualization [6] and database [83] communities agree that interactive data exploration usually does not have a clear, well-formed goal. Alspaugh et al. [6] distinguish exploration from analysis, the latter being when clear goals are formed. Battle and Heer [13] observe that exploratory visual analysis include a spectrum of goals specifications, from no goals at all, to clear a priory goals and/or hypotheses.

To the best of our knowledge, no work or tool in the database or visualization communities addresses the need to specify analysis or exploration goals or to support (let alone solve) data exploration through goal specification. However, other domains like web search [74] or exploratory search [71, 140, 176] have been trying to identify user's goal, especially by analyzing past actions. Notably, a clear categorization of the goal, like the one done in web search [25, 172], is still missing.

Analytical question. Concretely, Battle and Heer report in [13] the consensus that analysts decompose their analyses into smaller tasks that may be re-used across data sets, including understanding data semantics, characterizing data relationships, analyzing

causal relationships or hypothesis formulation and verification. This approach underscores the presence of the concept of "analytical questions" within the data narrative, where the goal of exploration is shaped by these questions. However, only a few works in the literature precise the semantics of these analytical tasks, and propose the formulation of the analytical questions or hypotheses behind them.

Two levels of abstraction addressing the formulation of an analytical question reflecting the intention of the data narrator were identified: (i) the intentions, addressed in [163] in a high-level language within the context of data cube exploration, from which collectors can be automatically derived; and, (ii) the intentions in a SQL-like language operating at a lower level of abstraction. SQL query addressed in [22, 154, 174] deal with analytical question but with a lower level of abstraction. In more details, Wongsuphasawat et al. [174] introduce compassQL, a visualization query language that consists of (1) a partial chart specification and enumeration constraints, and (2) directives for grouping, choosing, and ranking visualization suggestions. Bozzon et al. [22] introduced SeCoQL, a SQL-like language and protocol specifically designed for supporting exploratory search over data sources, translating user interactions. Singh et al. [154] support data exploration by introducing a SQL extension to express exploratory queries over databases.

Message. The essence of a data narrative lies in its conveyed messages, which are derived from the significant discoveries within the data. Interestingly, this concept was largely absent from the literature, with few exceptions. For instance, Hullman et al. [81] addressed the challenge of effectively communicating unfamiliar measurements to the audience. They developed tools to automatically reframe unfamiliar measurements by using the measurements of familiar objects.

Leppanen et al. [99], on the other hand, delved into the challenges surrounding automated journalism, exploring algorithms capable of transforming data into narrative news texts without human intervention.

Character. Characters are the main element of narrative in narratology domain [32, 49, 112]. Character refers to participants who play a role in the unfolding events described in the narrative. Metoyer et al. [116] proposed an approach to automatically extract character and the other narrative components from domain specific texts (sport game recap in their case) using natural language processing.

Measure. The central role that data plays in the narrative, is underscored by the diversity of elements it encompasses, including aggregates, patterns, outliers, relationships, correlations, and differences [170]. These numerical data, act as pivotal elements inspiring the narrative's message. Numerical data represents a quantitative information in diverse contexts [150] or a specific metric, often related to the core values of interest in the dataset named measure in data cube environment [163].

Shi et al. [150] identified 10 types of measures such as values, difference, proportion, trend, categorization, distribution, rank, association, extreme, and outlier. Lavalle et al. [97] categorized measures according to the level of measurements addressed in [108] (nominal, ordinal, ratio, and interval).

2.2.3 Form of expression: Structure

Messages gathered from data analysis need to be assembled into a plot that is interesting, illuminating, and compelling [98]. Several studies have dedicated their attention to the challenges associated with organizing a collection of messages by determining the most effective structure and appropriate ordering. In order to successfully address this challenge, it is necessary to consider two key aspects: the *structure* that refers to the overall organization and arrangement of a set of messages within a data narrative, and the *ordering* that pertains to the sequence in which the messages are presented within each part of the narrative. By carefully managing both structure and ordering, one can create a coherent and meaningful framework that takes into consideration the organization of the message set.

The fact of ordering and structuring messages is addressed by the information visualization and data journalism communities. The information visualization community abounds with studies aiming to understand how narrators structure visual narrations and propose strategies for organizing messages.

Structuring The structure of a narrative refers to the overall organization that manages the arrangement of the story's elements. It involves how the narrative is divided and organized to create a cohesive and meaningful whole. There are contributions in two notable directions: the *style of writing* and *level of interaction*. Some contributions for structuring a data narrative are depicted in Figure 2.2.

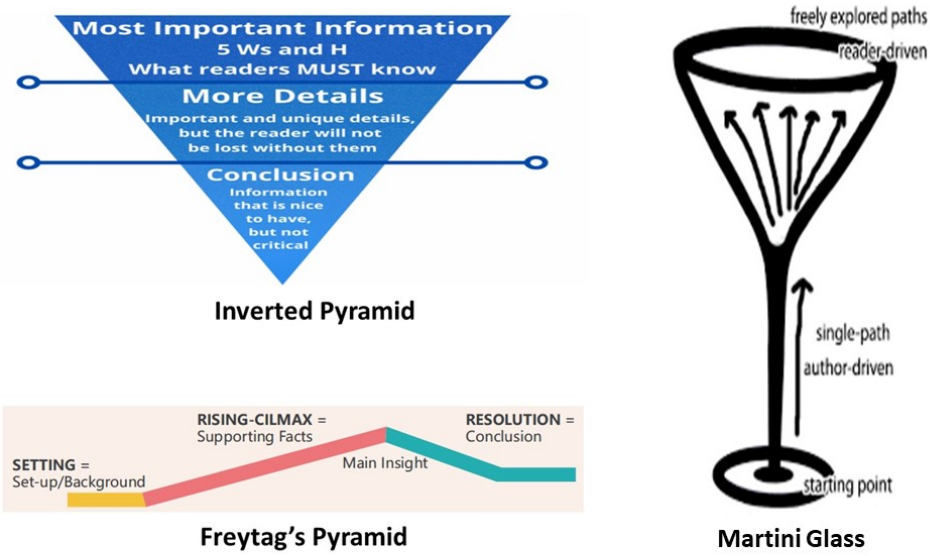


Figure 2.2: Some contributions for structuring data narratives

When it comes to writing style, data narratives can benefit from the traditional structures frequently used in fiction. A good example is the Freytag's pyramid [175] that organizes the story in three acts: the *setting*, that introduces background and main con-

cepts, the rising *climax*, where insights are chained guiding to the main insight, and the *resolution*, concluding the story. As another example, Gkesoulis et al. [68] approach the organization of data narratives through a cinematic lens. First, an introductory act is built with the initial query, and two subsequent acts are used to put context. These acts contain visualizations highlighting important facts, as well as text and audio describing these facts. A summary act concludes with all the important highlights of the previous acts.

Other styles come from practice, specifically from data journalism. In particular, Kosara et al. [91] claim that most data narratives follow the inverted pyramid style, a popular style of writing characterized by a clearly-defined and structured beginning without having a defined end. However, Kosara proposed a new pattern of structuring the data narrative, inspired from classical narrative, that starts out by a question or a claim (without revealing all insights up front), then provides evidences, and finally concludes by tying the evidence back to the initial claim or question.

When it comes to the level of interaction, researchers have developed diverse approaches for structuring a data narrative [147, 170]: a purely author-driven approach has a strict linear path, relies heavily on messages enriched with explanations and includes no interactivity, a purely reader-driven approach has no prescribed ordering of messages, no explanations, and a high degree of interactivity, while a hybrid approach mixes both strategies. Among hybrid approaches, one notable example is the Martini glass model. In this model, a predefined sequence of messages serves as the starting point, but thereafter, multiple navigation paths are available. This allows readers to delve deeper based on their individual preferences or specific requirements. They have the flexibility to explore different avenues according to their needs or personal interests.

Ordering The ordering of messages in a data narrative is vital for guiding the audience's comprehension. Various approaches have been explored to determine the most effective way to order the messages within a data narrative. These approaches include transitions, timeline designs, and the *application of deep learning techniques* using the set-to-sequence framework. Also, we studied how automatic data narrative tools order their data narratives.

Transitions are one aspect of ordering in a data narrative that focuses on the smooth progression between messages. Hullman et al. [80] identified possible transitions among visualization types, studied the cognitive cost of such transitions from the audience perspective and proposed to arrange visualizations in order to minimize the overall transition cost. Hullman et al. [82] addressed the ordering of narrative with visualizations to effectively express the intended narrative and identified two types of ordering: hierarchical ordering (grouping subsets of visualizations with shared data properties, such as a common measure, time period, spatial region, or level of aggregation) and parallel ordering (repeating a pattern of transitions two or more times in a sequence). In practice, in a survey of success infographics, Wang et al [168] found three types of ordering that are frequently used: (i) more than half of them followed random order, (ii) a quarter of them organized messages sequentially, and (iii) the remaining used multiple series, generally comparing two sequences.

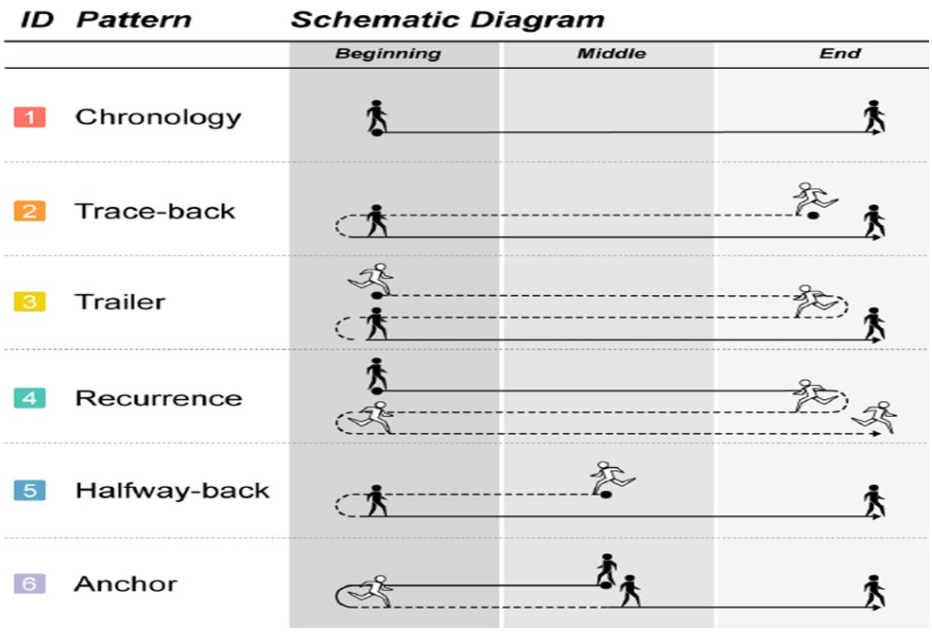


Figure 2.3: Patterns time-oriented stories from [96]

In another study, Lan et al. [96] analyzed a corpus of 80 time-oriented stories and identifed six most salient patterns of narrative orders represented in Figure 2.3: chronological and non-chronological order. They also conducted a crowdsourcing study with 221 participants and found that narratives with non-chronological order have the potential to make time-oriented stories more expressive without hindering comprehensibility.

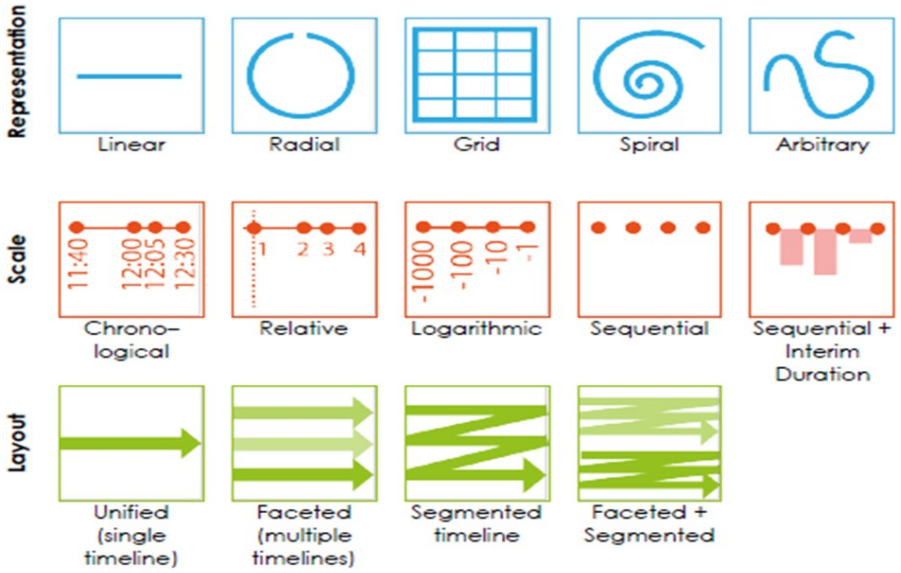


Figure 2.4: Timeline designs from [24]

Timeline designs, depicted in Figure 2.4, are another direction for ordering data narrative messages [24]. The authors identify 14 ordering choices which are further charac-

terized by: representation (the guiding visual metaphor; e.g., linear or grid), scale (the correspondence between temporal distances and distances on the display; e.g., chronological or relative), and layout (how the timeline is partitioned into separate regions of the display; e.g., unified in a single timeline or faceted into multiple timelines). In addition, Walker et al. [165] proposed to use story boarding techniques for visual analysis to organize the plot of the data narrative with the aim to explore different trending themes, explain, understand and presents events as they unfold.

Deep learning techniques are another direction for ordering data narrative messages [37, 66, 101] Logeswaran et al. [101] presented novel deep learning techniques utilizing the set-to-sequence framework to tackle the issue of effectively organizing a collection of sentences in a coherent and logical manner. This is a challenging task because of the difficulties in preserving the meaning and context of each sentence while ensuring that they are organized in a sensible order.

Regarding the ordering suggested by automatic data narrative tools, Zhang et al. [102] grouped messages based on their commonness, considering shared behaviors, as well as exceptions that deviated from the shared behavior. Similarly, Wang et al. [168] grouped messages that pertain to the same data, and a subset of messages was selected to maintain a balance between interestingness and diversity. Messages are initially presented in a random order and may sometimes be rearranged during visualization to fit the layout better. Another approach to ordering is employed by Calliope and Erato [150, 156], where messages are organized based on a tree structure. In this structure, nodes represent individual messages, while edges indicate transitions between them. The data narrative is then represented by a path within this tree, capturing the flow of information and storytelling.

Act. Several works concerning crafting movies broke down movies into simple acts, with each act serving as a container for grouping several messages. In [68], authors organize story plots in 3 acts with a clear semantics related to analytical questions: the first one provides contextualization for the characters as well as the incident that sets the story on the move, the second one build up protagonists actions and reactions and finally, the latter concludes.

Episode. Episodes play an important role in both the realm of cinema. They serve as the backbone for structuring the narrative, enabling an exploration of data-related subjects, Within a larger dramatic work, an episode is a cohesive narrative unit comprising a collection of events that occur relatively close to each other in a given partial order [106]. For instance, Stopler et al. [155] proposed techniques to communicate information in a creative way, which can help in accomplishing a purposeful episode. Several works [40, 100, 123] from text prediction domain can be used to improve the episode writing.

2.2.4 Substance of expression: Presentation and data visualization

Data visualization involves transforming complex data into visual representations, including charts, graphs, and other graphical elements, to enhance comprehension and interpretation. By converting raw data into visually appealing and easily understandable

forms, data visualization can effectively communicate messages, enabling more informed decision-making. In particular, data visualization can be leveraged to create compelling data narratives that convey complex information in an intuitive and impactful way, helping to clarify concepts and ideas for audiences.More concretely, data visualization plays a key role in preparing the substance of the discourse that includes the set of all media used to show structured messages. We organized the existing literature based on the characterizations of visual narratives, which refers to the use of visual elements. Additionally, we explored the definitions of dashboards, which are widely recognized as a popular form of data visualization comprising multiple components for effective information communication. Furthermore, we investigated the various components typically found within dashboards.

Visual narrative. While several characterizations of visual narrative exist, many of them address nuances that can hardly fit into a single and comprehensive definition [9]. This also happens since characterization is often driven by partially overlapping sets of papers, where authors focus on specific narrative patterns (e.g., [80, 114, 147, 155]). Thus, we refer to a coarser characterization following well-agreed coordinates, such as *forms* [34, 147], and *interaction* [80, 151].

Several works [34, 147] have categorized numerous types of visual narratives.The authors propose a classification scheme, depicted in Figure 2.5, that identifies and highlights seven categories based on their narrative degree, complexity, and explanability. These categories include dashboard, annotated chart, infographic, scrollytelling, data cominc, timeline or storyline, and data video.

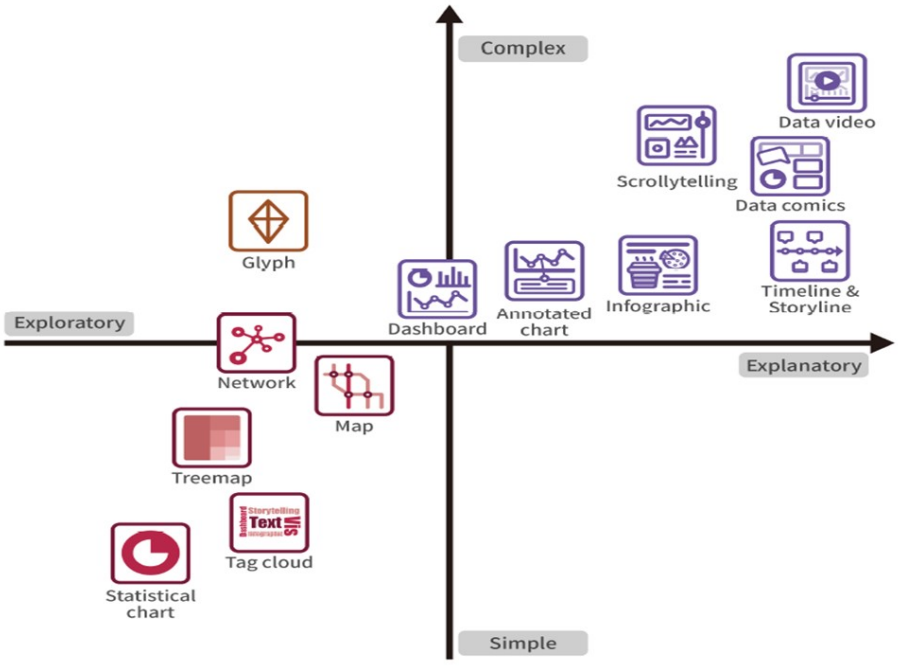


Figure 2.5: Visualization forms from [34]

Interaction identifies the primitives to manipulate and get the most out of such data visualization [76]; indeed, since exploring information collections becomes increasingly difficult as the volume grows, it may be difficult to locate known items or to browse to

gain an overview. Effective narratives should also choose the proper level accordingly to well known visual information mantra "Overview first, zoom and filter, then details-on-demand" [151]: gain an overview of the entire collection, zoom in on items of interest and filter out uninteresting items, finally, select an item or group and get details when needed.

Dashboard for data narrative. A dashboard is a well-known design that includes various components for the visual communication of information. Dashboards comprise a collection of data visualizations, with these visual elements forming the essential constituents of the dashboard. These components are designed to present data in a concise and meaningful way, allowing users to easily interpret and analyze complex information. Each dashboard component is typically tailored to display a specific type of data, such as charts, graphs, or tables, and is organized in a logical manner to facilitate the user's understanding of the underlying data.

According to [57, 105, 113, 171, 178, 178], dashboard is as an interactive performance management tool that provides employees with timely personalized information to enable them to monitor and analyze the performance of the organization. Dashboards usually group multiple tables and charts, enabling users to make decisions based on data-driven insights [97], and act as cognitive tools that help to visually identify trends, patterns, and anomalies [23, 57].

Effective dashboards are the product of informed design and several works have been done in the field of data visualization to create and design a dashboard. Once the data sources, metrics, and relationship are defined, dashboard components can be either manually organized to support specific types of cognitive tasks [19] or automatically derived [73, 97] given the nature of the data to be displayed.

Dashboard components for data narrative. Dashboard components aid data access and emphasize key information [60]. To pick the proper components the following questions should be kept in mind [23]: "what KPIs should data users see?", "what is the proper visual representation for the given KPIs?", and "do the KPIs need additional context?". Unfortunately, choosing the wrong components or simply defaulting to the most common type of charts (e.g., picking a line chart over a column chart when discrete categorical data are in place) could confuse the viewer or lead to mistaken data interpretation.

A simple guide [73] had been proposed to help data narrators in choosing the right components, for instance, through visual properties such as position, size, shape, and color (e.g., as in [33, 76]). In [158], the authors map visual and narrative elements which help the data narrator in choosing what visual elements will use to convey the most essential element of the narrative.

2.3 Data narrative crafting

In this section, we review the works describing the internals of the data narration process, as well as the tools that automate (part of) the crafting process.

2.3.1 Data narration processes

As previously discussed, data narration is a complex process, at the crossroads of several domains: data exploration, data visualization, data management, etc. Despite the many contributions in each of these areas, few works offer comprehensive workflows describing the entire data narration process. In this context, we will outline the main phases that characterize the process and delve into the activities involved in crafting a data narrative.

Process There is general agreement that the process of data narrative crafting can be divided into three basic phases, as shown in a number of studies [35, 92, 98].

Firstly, the phase of analyzing and exploring data, which involves examining various aspects of complex data and investigating the interrelationships between them. Secondly, the phase of structuring and organizing messages, during which messages are arranged to construct a compelling narrative. Lastly, the phase of presenting, which entails visualizing and materializing the structured messages in a visual format to effectively convey the information to the audience. Several patterns for crafting data narratives have been identified [47, 98, 149, 169] and presented in Figure 2.6. Some approaches involve only two phases: analyzing and presenting the data. On the other hand, other methods emphasize the importance of three phases: analyzing, structuring, and presenting the data. Within these three phases, the order of execution and the interplay between them are distinguished and further elaborated upon.

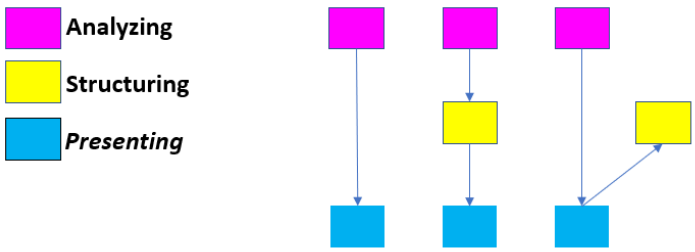


Figure 2.6: Patterns for data narrative crafting

Duangphummet et al. [47] delved deeper into data narrative crafting and proposed a protocol consisting of five phases. These phases align with the previous approaches but offer additional insights and elaboration, providing a more comprehensive and detailed framework for creating data narratives: *conceptualization* of the data narrative domain, targeted audience and distribution channel, *data preparation* to deliver data that is relevant to the use, *realization* to deliver a storyline with detailed content and an initial form of key visualizations, *visualization design* to redesign the visualizations and create visualization prototypes, and finally, the *visualization development* where technical requirements are defined, and the key visualizations for target devices are developed and deployed.

Apart from the protocol described previously, a detailed process of crafting data videos is described by Shi et al. [149]. It consists of four phases with a different ordering with a distinct ordering where the structuring phase comes after the presenting phase: (i) *collecting a series of data facts* around a certain topic, (ii) *constructing a storyline* as an assembly of these data facts into a sequence, (iii) *choosing data visualizations* for the

data facts and deciding how to animate them by drawing a storyboard, and finally, (iv) *realizing the storyboard* via a design software in which the narrator edits and combines the animated visualizations until a coherent data video is accomplished.

Many works [98, 169] underline the importance of moving between the data narrative crafting phases and in particularly, the need to *move back and forth between the narrative, visualizations, and the data.*

Activities Besides the previously described works proposing global crafting processes, some works describe the necessary activities to be conducted. Battle and Heer [13] identified three ways to start a data narrative: having a precise idea in mind, having a vague idea refined during data exploration, or having no idea before exploring the data. In particular, Weber et al. point that the crafting process starts by either an idea, a problem or a question [170]. This aligns with the observations made by Battle and Heer [13], who identified three distinct approaches to creating a data narrative: having a precise idea in mind, having a vague idea refined during data exploration, or having no idea before exploring the data.

Notably, many works deal with the phase of structuring the narrative [149, 168] and underline the importance of different story structures and different kinds of interactivity in data narration [147, 170].

Finally, very few works highlight the importance of intentional aspects. Bach et al. [9] found 18 narrative patterns to provide guidance on how to achieve five general storytelling intents (i.e., argumentation, flow, framing, emotion, and engagement). Similarly, design patterns have been proposed for data comics [10]. Thudt et al. [159] stress that subjective perspectives can be introduced at every step of visualization creation: during data collection and processing, visual encoding, and when refining the presentation.

2.3.2 Automated data narration

In recent years, there has been a growing body of research focused on automating the generation of data narratives, which has shed light on how the process of data narrative crafting is approached. Some systems are depicted in Figure 2.7.

These works [68, 132, 149, 150, 156, 168] have aligned with the previously identified phases, namely the phase of analyzing and exploring data, the phase of structuring and organizing messages, and the phase of visual presenting and communicating messages. However, it is important to note that in certain tools or approaches like Calliope [150] (see Figure 2.7), there may not be a clear distinction between the structuring and analysis phases. The absence of this distinction can be attributed to the interconnected nature of these two phases, as analyzing the data often involves making decisions about how to structure the narrative based on the messages crafted.

Subsequently, we will provide a comprehensive description of the data narrative crafting tools.

In particular, Wang et al. [168] conducted a qualitative analysis of 245 infographics examples to explore the infographics design space in terms of structures, sheet layouts, fact[1] types, and visualization styles. Based on those, the authors propose Datashot, a

[1]It is worth noting that the term *fact*, used by many authors of the visualization community, corresponds to the concept of *finding* discussed Subsection 2.2.

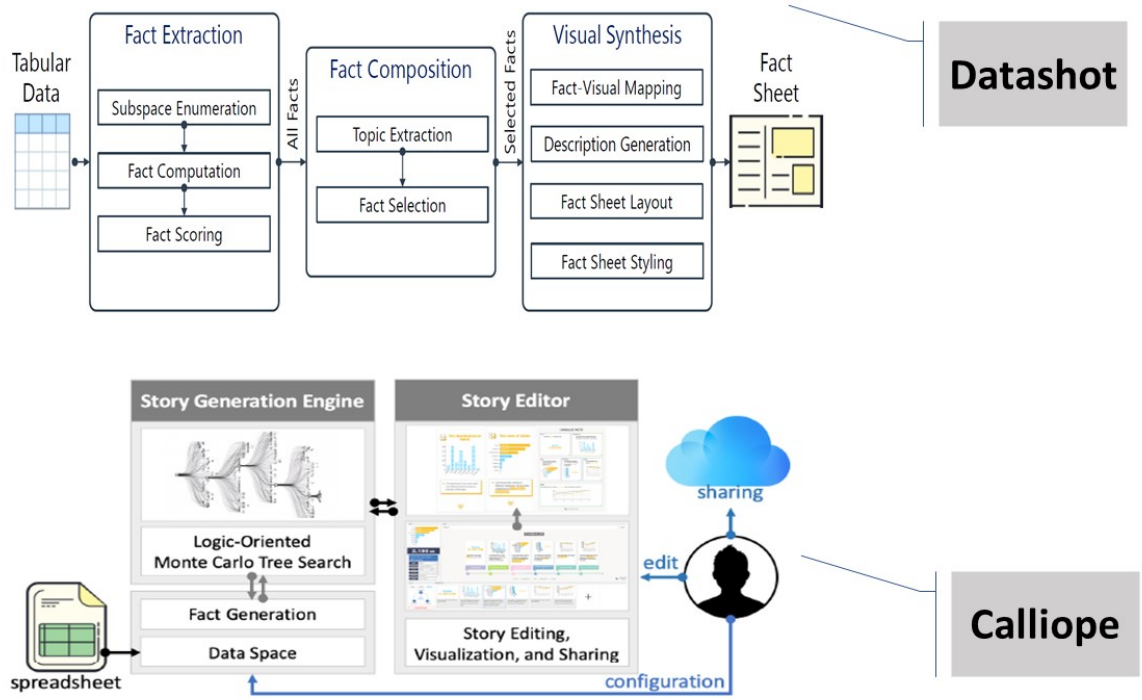


Figure 2.7: Datashot [168] and Calliope [150] systems overview for data narration

system for supporting a fact sheet generation pipeline consisting of three phases: (i) fact extraction, (ii) fact composition, and (iii) presentation synthesis.

While Datashot follows a three-phase process, Calliope and Erato are two-phase systems. In more details, Shi et al. [150] proposed Calliope, a system that can automatically generate visual data stories with facts arranged into a logical sequence. It consists of two main modules: (i) the story generation engine, for generating, choosing and organizing the facts that will participate in the narrative, and (ii) the story editor, that visualizes the data story (generated as a series of visualization charts) and allows the users to change it based on their preferences.

Sun et al. [156] proposed Erato, a human-machine cooperative data story editing system that allows users to generate data stories. It consists of three major modules: (i) a fact embedder that takes a fact's specification string as the input and converts it into a vector representation, (ii) an interpolator that generates new story content by interpolating between two data facts, and (iii) a story editor that enables user to verify, refine, and incorporate data facts to make a more smooth and compelling story.

Similar to Calliope system, user interaction is considered as an integral part of their systems. Storyfacets [132], is a system that focuses on the capture of analysis actions as abstract provenance operations. Further, the system visualizes the analysis history using multiple visual formats to facilitate collaboration among different user types. The workflow of Storyfacets begins with analysts exploring data in the trail view, which provides a comprehensive analysis interface. In the background, the system automatically maintains multiple views, including the trail view, dashboard view, and slideshow view, catering to different user preferences and interaction styles.

Unlike the previous tools, Gkesoulis et al. [68] follow a structured approach based on

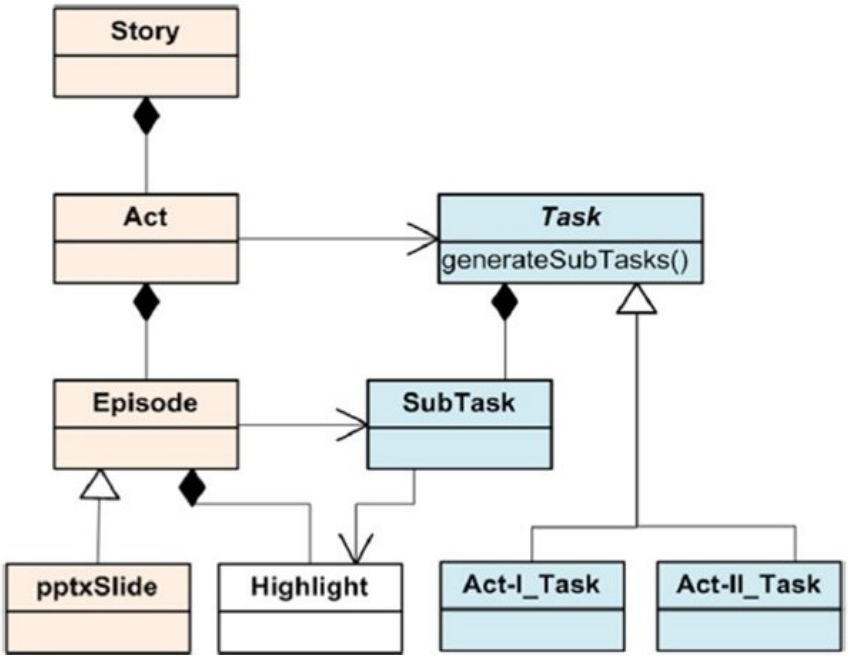


Figure 2.8: Mechanism for Cinecubes [68]

the concept of acts, scenes, and episodes. The authors detail the process of crafting a data movie in the form of a powerpoint presentation, to answer a specific user's need described by a query. Such approach, used for the Cinecubes is represented in Figure 2.8. CineCubes system follows a structured approach consisting of approximately three acts, resembling a movie narrative, to present data in a concise and engaging manner through PowerPoint slides, with each slide representing an episode and containing relevant queries and visualizations to convey insights to users. The first act sets the context and introduces the incident that drives the story forward, the second act builds up the actions and reactions of the characters, and the third act resolves the film's plot. Each act is further divided into sequences of scenes, with each scene bringing about a change in the plot's status. The data narrative is organized into episodes, which practically correspond to slides in a PowerPoint presentation. The structure is result-based, with each episode accompanied by a procedural-based approach. The generation of queries and slides within each act is handled by an abstract class called task. Each episode contains one or more queries in the background, and these queries are executed by subtasks. Each subtask is responsible for gathering specific data, which is then visualized in the main part of the slide. Additionally, the class 'highlights' is responsible for determining key findings within each episode.

Take away. The process of crafting data narratives is generally categorized into three main phases serving for exploring data, structuring messages crafted on the basis of exploring data, and presenting visually the structured messages.

Concerning the automatic tools for generating data narratives, the construction is linear in the sense that there is no back and forth movements between phases. In addition, they target a specific domain or data format and organize stories according to pre-established patterns. In particular, we highlight the absence of intentions, that are,

at best, modeled via an initial query or a topic.

2.5 Lessons learned

In order to understand the domain of data narrative, we summarized the reviewed literature into answer of these questions: What is a data narrative? What are the component of data narrative? How does one create a data narrative?

By structuring our discussion around these key questions, we will develop a comprehensive understanding of the state of the art of the data narrative. Hereafter, a summary of lessons learned.

- **What is a data narrative?** Data narrative is a set of messages supported by data. Messages are retrieved from data analysis and are structured in a certain order and communicated visually to the audience.

- **What are the concepts of data narrative?** Table 2.1 summarizes the main concepts of data narrative retrieved from the literature. The fact of exploring data through asking analytical questions and implementing collectors can lead to the discovery of various findings, which are subsequently communicated to the intended audience in the form of messages. To effectively communicate these messages, they are often presented in a structured and organized manner, such as through a compelling plot of data narrative formed by a classical representation acts and episodes. Additionally, it is common practice to analyze and interpret these findings within their appropriate context before disseminating them to the intended audience.

  Observations It is important to note that the existing literature presents a preliminary version of the concept of a message. This initial version primarily focuses on the communicating of findings derived from data, without fully capturing the intentions of the data narrator. Furthermore, there is a lack of comprehensive information regarding the form and composition of the message, including its description and specific details. It is worth mentioning that the notion of user models and human in the loop has emerged as a potential approach to address this problem [36, 148]. By involving users more actively in data narration, a more effective approach to crafting meaningful messages from data can be achieved. However, we also noticed a lack of attention given to the intention aspect in the literature. This means the substance of the story, i.e., the composition of story elements (hypothesis, messages, etc.) as pre-processed by the author's cultural code [32] is ignored. We claim that this absence is regrettable; if data narrations are to be shared, reused, and have their crafting process documented, then the intentional aspects deserve more attention.

- **How does one create a data narrative?**

  Most of the works describing the data narration process agree on the three general phases:

  - The phase of analyzing data refers to retrieve findings among data,

Concepts	Summary	References
Exploration	refers to the act of investigating and analyzing data to uncover patterns, relationships, and insights that can inform decision making	[20, 55, 86] [136–138] [15, 29, 50] [67, 72, 83]
Collector	refers to the tool, method or simple query used to request or gather data from various sources	[7, 118]
Finding	is a significant discovery that emerges from the analysis	[2, 64, 107, 143, 144] [17, 35, 75, 157, 179] [67, 84, 149, 150, 168]
Goal	refers to the main objective of a data narrative	[6, 83] [13, 71, 74, 140] [25, 172, 176]
Analytical question	reflects intention of the data narrator and represents a particular aspect of the objective	[13, 163, 174] [22, 154]
Message	conveys only interesting and/or important findings	[81, 99]
Character	refers to the qualitative aspects of a message and captures the elements that give a message its flavor	[32, 49, 112, 116]
Measure	refers to the quantitative aspects of a message and captures the objective elements that can be counted	[150, 163, 170]
Plot	is an arrangement of messages in a way easily understandable by the audience	[68, 91, 98, 170, 175] [80, 82, 147, 168] [24, 37, 101, 165] [66, 102, 150, 168]
Act	constituent part of the plot	[68]
Episode	granular piece of the narrative that conveys a message	[40, 100, 106, 123, 155]
Visual narrative	refers to the plot that is conveyed through visual means	[9, 80, 147, 155] [24, 76, 80, 114, 151]
Dashboard	is a visual interface that presents important information	[57, 105, 113, 171, 178, 178] [23, 57, 97, 135] [19, 73, 97, 135]
Dashboard component	is an individual element that make up a dashboard	[23, 33, 60, 73, 76, 158]

Table 2.1: Main concepts of data narrative identified from the literature

– The phase of structuring messages refers to organize the messages crafted into

narrative pieces,

– The phase of presenting messages refers to present the structured messages through crafting visual artifacts.

Observations We remarked that a significant amount of research has been devoted to exploring and automating the process of data exploration, as well as the visual aspect of data narration, which is closely linked to the data visualization community. However, the automated data narration is still in its infancy, mainly applying rigid patterns and lacking the necessary flexibility of moving between the three phases.

Chapter 3

A conceptual model for data narratives

Albert Einstein said: '*Any fool can know. The point is to understand*'. In line with Einstein's insightful quote, the importance of understanding, rather than mere knowledge, is crucial in the context of data narratives. The domain of data narratives can be either familiar or unfamiliar to readers, with some having a basic knowledge of the term derived from the combination of "data" and "narratives". However, there are readers who may still lack a comprehensive grasp of the concept.

The goal of this chapter is to bridge that gap and provide readers with a comprehensive understanding of the core concepts of data narratives. Drawing inspiration from Einstein's quote, we aim to present the domain of data narratives in a straightforward and accessible manner. Our mission is to ensure that readers do not feel overwhelmed, but instead gain a clear understanding of the data narrative domain.
By achieving this, we establish a conceptual model for data narratives, which facilitates understanding and encourages the reuse of data narratives in various domains, such as data journalism, business intelligence, and scientific research.

Contents

3.1 Introduction

This chapter focuses on conceptualizing the domain of data narrative in order to facilitate understanding, sharing, and reuse of data narratives, considering the lack of existing data narrative modeling and its main concepts as discussed in previous chapter. We propose a novel 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. The primary objective of this model is to identify and articulate the fundamental concepts that back up a data narrative across different domains, specifically in the realms of data exploration, data visualization, and narration. By doing so, the model helps data narrators in comprehending and preparing the essential components required to create a compelling data narrative.

This chapter aims to introduce data narrative domain through the presentation of the definition of data narratives and a motivating example. Subsequently, we present a novel conceptual model in detail, offering a structured and principled definition of the key concepts within the domain. This model unfolds the relationships between these concepts, clarifies their roles and usage, and primarily assists authors in conceptualizing and constructing compelling data narratives. Furthermore, we present the experiments realized to prove the feasibility of the proposed model through an approach for reverse engineering a data narrative from the analysis it's based on and a proof of concept implemented in the form of a web application.

Moreover, we list a compilation of works, which offers direct actionable contributions for implementing the concepts of the data narrative model. Finally, the chapter concludes by presenting important summaries of the discussed content.

It should be noted that both the conceptual model for data narratives [128] and the proof of concept [127] were published at the ER conference in 2020, including a demo. A french version of the conceptual model was published in EDA [129].

3.2 Model overview

This section aims to offer an insightful overview of the data narrative model. Firstly, we present an example of a data narrative, carefully illustrating the practical application of the model's concepts. Secondly, we propose a concise definition of data narratives, outlining the key characteristics of such data narrative.
Through this combination of definition and example, readers will gain a deeper understanding of the data narrative model and its relevance to understand and reuse a data narrative.

3.2.1 Motivating example

This subsection illustrates the components of our model (signaled in italics) using a simple visual data narrative about women and strokes, published by GOOD[1]. For illustration purpose, we describe a plausible process for defining analytical questions and collecting data, which is not precised by the author. The final result, a *visual narrative* is depicted

[1] https://www.good.is/infographics/facts-about-women-and-strokes

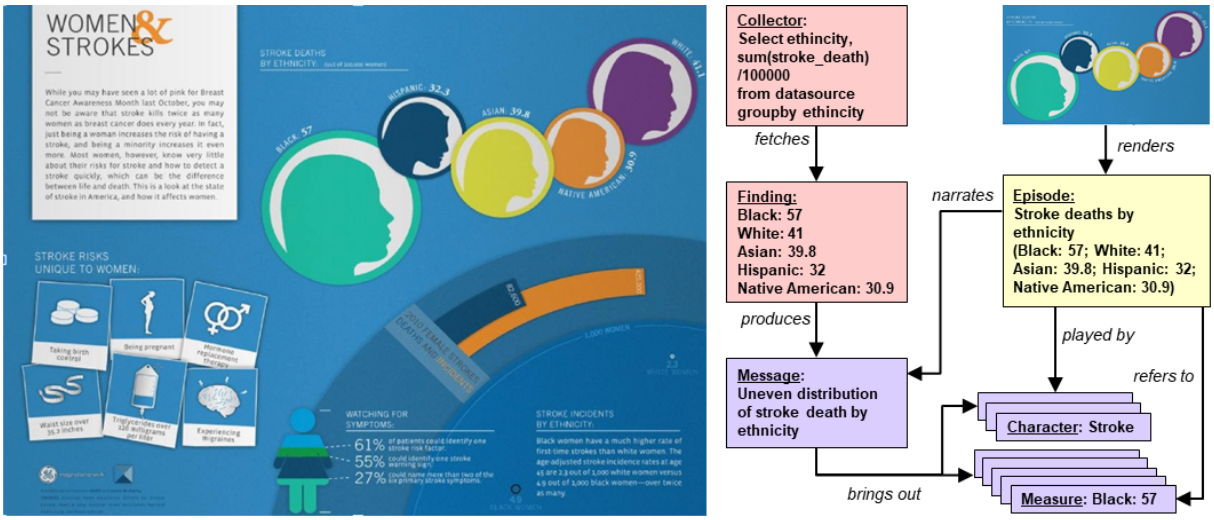


Figure 3.1: Example of data narrative, and a partial object diagram for a particular message (right).

in Figure 3.1 (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 structures the discourse by arranging messages in coherent pieces of discourse: *acts* narrating a major piece of information and a major part of the narration with a significant communication, and *episodes*, subparts of lesser importance on their own, narrating specific messages. In this example, there is a unique act and six episodes. 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 Figure 3.1 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 author's intention. Here, the author'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 or concept of the story, in addition to women and stroke, both already pointed as characters by the analysis goal.

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 author, who called several *collectors* for analysing data and collecting *findings* in order to answer analytical questions.
For example, a collector queried 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).

3.2.2 Data narrative definition

After examining the narratives and data narratives presented in Section 2.1, the following definition for data narrative, which we believe captures both its core and extent, is proposed.

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 borrow Chatman's terminology and extend his structure of narrative considering that data narrative must describe how the content of the story (Chatman's events and existents) is derived from data.

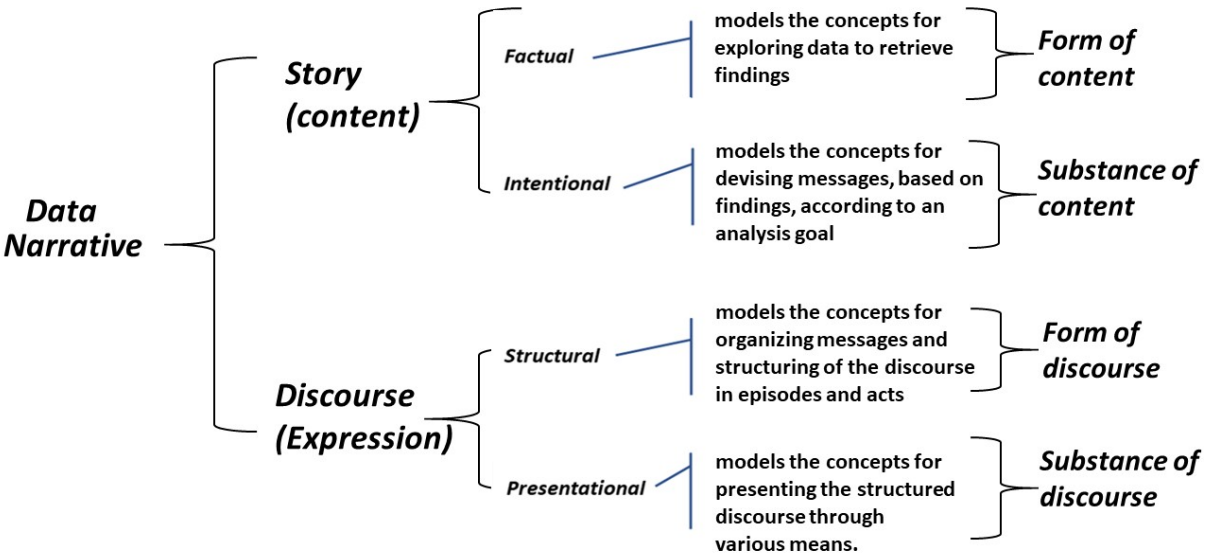


Figure 3.2: From narrative to data narrative

As illustrated by Figure 3.2, we distinguished 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. Specifically, and originally compared to classical models of narration, the factual layer includes an entity for *findings* based on facts and models collected from data, and the intentional layer includes entities for *messages* derived from findings, where the narrative *characters* (e.g., important business objects) demonstrate interesting measurements. 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.
Note that our model of data narrative is agnostic of a specific data model; all the specific details on how data and facts are produced to serve the information goal and support the extraction of findings are encapsulated in a *collector* entity.
The structural layer includes entities modeling the arrangement of the messages into a structured combination of presentable discourse elements, and the presentational layer includes entities for the assignment of a presentable set of media to each of the narrative's discourse structure.
Figure 3.1 (right side) illustrates a partial object diagram concerning message m_2, from the collection of findings to the rendering of an episode.

3.3 The model

This section presents a conceptual model for data narratives. This model is depicted in Figure 3.3, organized in four layers and expressed using UML class diagram notation. We omitted class properties in the model for readability purposes.
The organization in 4 layers, adapted from Chatman [32], reflects the transition from raw facts to the visuals communicated to the audience of the data narrative. On their way to the reader, the facts traverse:

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. *Findings* emerged 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* the author 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*.

4. **Presentational layer**. The presentational layer models the rendering of the data narrative, i.e., a *visual narrative*, that is communicated to the reader through visual artifacts (*dashboards* and *dashboard components*).

The identification of key concepts in the model is the result of a comprehensive analysis of the state of the art, as outlined in Chapter 2. These essential concepts are documented in Table 2.1, capturing their significance in the context of this research.

To understand the organization of the model, one should note 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 in the data that answer a, and may raise new, analytical question(s). This specific message is then the discourse structural building block: episodes narrate specific messages, and acts, built as sets of episodes, narrate a broader message. A message is also indirectly connected to the presentational layer: a global message is visually conveyed by one dashboard, each of the dashboard components illustrating one specific message.

In what follows, we will delve into an in-depth study of each layer, examining its individual components and interconnections in detail. Additionally, we will revisit the example discussed in Subsection 3.2.1 to provide concrete illustrations of the concept.

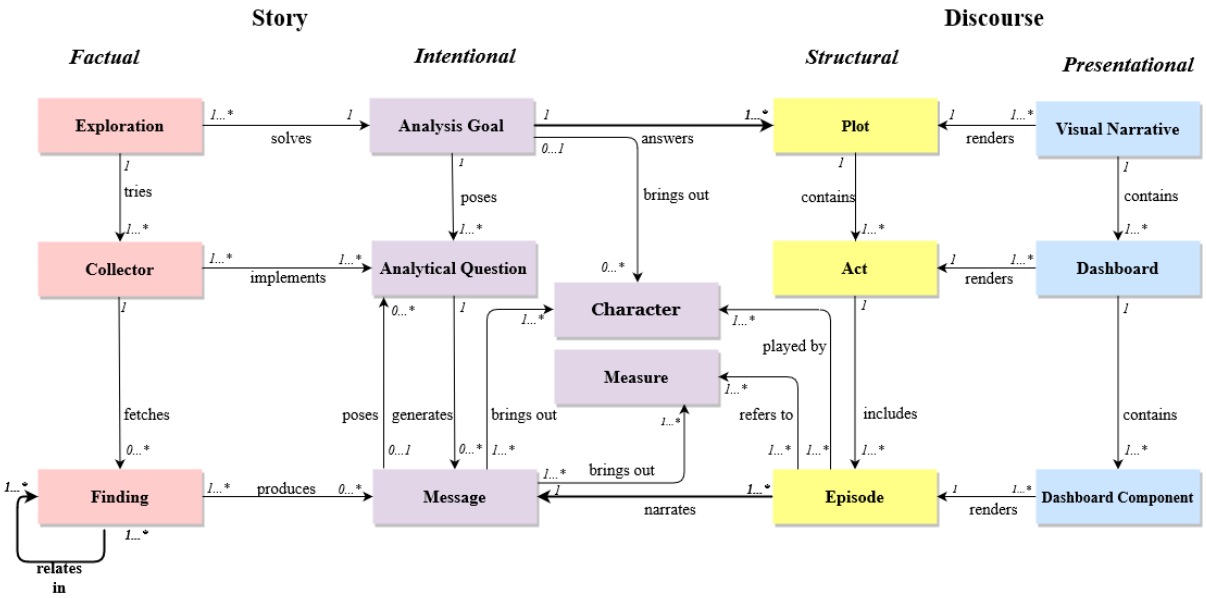


Figure 3.3: The data narrative model, organized in layers (relations in bold were extended from the version published in [128])

3.3.1 Factual layer.

Data narratives need data. Data represent the facts that support a story. The factual layer concerns the fact of looking for data in a set of data sources, analyzing them, and obtaining added value and findings. Concepts of the factual layer are:

- An **exploration**, or data exploration, refers to the fact of investigating and analyzing data to uncover patterns, relationships, and insights that can inform decision-making. It involves a systematic and structured approach to analyzing data, using various techniques such as visualization, statistical analysis, and machine learning algorithms.
- A **collector** is an algorithm or simple query used to request or gather data from various sources. For example, a collector can be a simple SQL query as represented in Figure 3.1 (on the right side).

- A **finding** is a significant insight or discovery that emerges from the analysis of collected data. Findings are, among the facts retrieved by collectors, those that are more striking, surprising, or relevant and worth narrating in the story. A finding can be: (i) the result of the collector with some significance, or (ii) a visual representation, such as a chart, graph, or map, to make the conclusion more engaging and easier to understand. Findings are the key concept of the factual layer. As an illustration, a finding can be represented by the top 5 values obtained from an SQL query, as depicted in Figure 3.1 (on the right side).

Bridging concepts Data are collected via a set of *collectors*, that can be queries in a query language or interface supported by the data sources, extraction tools (e.g. wrappers or loaders), or more generally, all kind of programs able to interact and retrieve data from sources. They may just retrieve data from sources or include functionalities for filtering, checking, building models and reasoning with data. For example, a collector may cluster data, compute correlations, detect outliers or emphasize contradictory data in order to produce insights. For the sake of generality, we do not assume a particular structuring of data (e.g., databases or unstructured files) nor a particular way of collecting findings (e.g., via queries, data analytic or other algorithms), but use collectors as an abstraction of data access and manipulation. A collector may *fetch* one or several findings, or conversely, it may evidence no narrating-worthy finding. The set of actions conducted to collect findings is called a data *exploration*. It aims at keeping trace of the set of collectors tried for addressing an information need. Each collector is part of an exploration while an exploration typically *tries* many collectors.

3.3.2 Intentional layer.

The intentional layer models the devising of the substance of the story based upon the author's analysis goal. Concepts of the intentional layer are:

- An **analysis goal** represents the main objective of a story, i.e., the intended information that should be explained, detailed and transmitted to the reader. It represents the global topic or theme that the data narrative aims to address or explore. The goal may bring out characters involved in the narrative and provides a high-level direction for the analysis. As an illustration of the goal, the goal of the data narrative depicted in Figure 3.1 is to " narrate facts about women and strokes in the USA".

- An **analytical question** represents a particular aspect of the goal and reflects the data narrator's intent when crafting the data narrative. An analytical question is derived from the goal and focuses on specific aspects related to the goal. An analysis goal poses one or several questions. These questions are more specific and help to delve deeper into the subject matter. An example of an analytical question is " Which characteristics of women (age, ethnicity, weight, etc.) have an impact on stroke deaths? "

- A **message** is, at the same time, (a) a partial answer to the question asked by the data narrator, and (b) the distilling, merging, and translation of a set of related

findings into information that is to be conveyed to the audience. Messages are the key concept of the intentional layer. An example of message is "the uneven distribution of stroke death by ethnicity"

- A **character** captures important data values that characterize a message and their fact-based quantification. Precisely, character refers to the qualitative aspects of a message and captures the elements that give a message its flavor. An example of character can be the ethnicity i.e., a relevant entity or concept of the story, in addition to women and stroke, both already pointed as characters by the analysis goal.

- A **measure** also captures important data values that characterize a message and their fact-based quantification. Precisely, measure refers to the quantitative aspects of a message and captures the objective elements that can be measured or counted. As an illustration, the ratios by ethnicity are brought out as relevant *measures*, i.e., relevant figures in the story.

Bridging concepts A goal is carved up into a set of *analytical questions*, each one concerning specific aspects of the goal, and a set of *messages* are raised in order to answer these questions. Indeed, a goal *poses* a set of analytical questions, and an analytical question *generates* a set of messages, possibly none. Goals are deeply related to explorations. Sometimes, an exploration is built for solving a clear goal, other times, the goal is progressively shaped while exploring data, but frequently, there is an interactive process tying a goal and an exploration. The underlying idea is that an exploration *solves* a goal and several explorations can be devised for solving a goal. In the same way, an analytical question can be solved by one or more collectors, each collector providing *implementation* means to one or more analytical questions.

As the data narrator explores the data and new findings are collected, progressively the author distills and structures them in their mind.

The findings raised during the exploration *produce* messages for building the story. A finding may produce many messages and a message results from one or several findings. Specifically, a message is (i) the interpretation of the finding and (ii) the description of the finding, including a detailed analysis of the finding as well as a discussion of the finding's implications and recommended actions. Possibly, new analytical questions can be *posed* based on the message found, inviting for more exploration. To further structure messages, we introduce two important components of them, *characters* and *measures*, that capture important data values that characterize a message and their fact-based quantification. Both characters and measures belong to the universe understandable by the audience. Messages *bring out* characters (e.g., a set of products causing a drop in sales) and the related measures (e.g. amount of sales for those products). In addition, some characters are previously known and *brought out* directly by the analysis goal (e.g., sales in France).

While a finding can be the result of the visual representation of the collector with some significance like an association rule, a path in a decision tree, or a message, on the other hand, is the answer to the intentional question that exploits a finding to label a a character concerning other characters or a measure. The following examples illustrate the difference between these two concepts:

- By comparing *Daily Infection* in *France* to *EU Average*, we *find* that they are similar. A message corresponds to the labeling of measure *Daily Infections* of character *France* with respect to another peer character, *Europe*.

- The message that a *Media outlet* cannot determine the existence of fake news answers the following hypothesis (analytical question): can the outlet solely determine fake news? This message follows from the finding that, by correlating the character *News Authenticity* to the character *Media outlet*, we find a non-significant correlation.

Noticeably, messages are the corner stone for structuring the story: an episode *narrates* a message. In this way, the intentional layer acts as a bridge between the exploration of facts (factual layer) and the structuring of the story (structural layer). In particular, a message, based on a finding, is the basis for building an episode.

3.3.3 Structural layer

This layer concerns the form of expression of the data narrative. While previous layers deal with the contents of the narrative, this layer focuses on its discourse. It is important to stress, that there is a design part served here. The idea is to address a goal via analytical questions and exploring data, the analysis has resulted in a set of messages to be conveyed to the audience. As reported in Section 2.1, the literature suggests to present messages that, there is a synthesis of a story as a coherent composition, where messages must be conveyed to the audience in an organized way [32, 35, 92]. Plots, acts and episodes model the parts of the synthesis of the narrative's content into this composition. Concepts of the structural layer are:

- A **plot** is the arrangement of messages in a way easily understandable by the audience. As an illustration, the plot is the arrangement of six messages detailed previously in Subsection 3.2.1 in the form of one act. Plot=$\{act_1\}$

- An **act** is the constituent part of the plot, which is the mean to convey a specific piece of information. As an illustration, the plot of the data narrative consists of a single act, which is equivalent to the plot itself. Act=$\{m_1,m_2,...,m_6\}$

- An **episode** is the key concept of the structural layer and the granular piece of the narrative that conveys a message. An episode consists of annotations, descriptions, and comprehensive explanations of a message that will be narrated within the episode scope. As an illustration, an episode can be "the uneven distribution of stroke death by ethnicity, black: 57, white: 41,..."

Bridging concepts To arrange the messages into a coherent and understandable order, the author must put in order a part of the audience-intended report with the messages that conveys an interesting piece of information. Following the terminology of traditional narration, we introduce an *act* as a constituent part of the plot, which is a major piece of information and a major part of the plot. Each act is composed of several subparts of lesser importance on their own, which we call episodes. A plot *includes* one or several acts and an act *includes* one or several episodes. A plot *answers* a goal and an episode *narrates*

a message, the episode text being the shaping of the message. Accordingly, characters and measures brought out by the message appear in the episode text, possibly starring or being highlighted according to author's narration style. One or many characters can *play* in one or many episodes. Analogously, one or many measures can *be referred* in one or many episodes.

3.3.4 Presentational layer.

This layer focuses on how the structured story is presented to the audience. Acts and episodes are represented and organized in order to be understood by the audience. Visualization aspects are the focus of this layer. A *visual narrative renders* a plot. It can be a slideshow, a notebook, a blog, or any other visual art allowing for the visual representation of a story. Concepts of the presentational layer are:

- Visual narrative refers to the plot (arrangement of episodes) that is conveyed through visual means such as images, illustrations, or other visual media.
- A dashboard is a visual interface or display that presents important information or data in a condensed and easily understandable format.
- A dashboard component refers to the individual elements or widgets that make up a dashboard. These can include graphs, charts, tables, text boxes, and other visual or interactive elements that display relevant episode.

Bridging concepts A visual narrative *contains* a set of *dashboards*, each one *rendering* an act. We use the term dashboard since it is general enough to accommodate varied types of visualizations (e.g. a Business Intelligence dashboard, an infographics, a section in a python notebook, a section in a blog or web page). In the same way, a dashboard contains a set of *dashboard components*, each one *rendering* an episode. Dashboard component is the key concept in the presentational layer and can include text, images, charts, maps, animations, etc. We remark that a story can be rendered in several ways or formats (e.g., an infographics and a video). In the same way, acts and episodes can be rendered by several dashboards and dashboard components.

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. A specific message is rooted in the facts analyzed, conveying essential findings, potentially raising new analytical questions. The message allows introducing episodes, the building blocks of the discourse. Each episode of the discourse 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. A message can narrated via multiple episodes, and the way these episodes are arranged determines the plot of the data narrative. Varying the arrangement of episodes leads to different plots. It's important to note that the plot reflects the development of the analysis goal, and since an analysis goal can have multiple plots, there can be different ways of presenting the same messages.

A version of the proposed model was published in [128] and the present one is published in [130]. The extension was done after practical experience, analyzing and creating several data narratives. Specifically, the model is extended by adding one relation and modifying some cardinally between concepts.

3.4 Experiments

In this section, we aim to evaluate the practical applicability of the proposed conceptual model for data narratives. To achieve this, in Subsection 3.4.1, we conduct a manual analysis of a data narrative and its associated analysis support, including the exploration and its results, the objective, and the analytical questions raised during the data narrative crafting, in order to confirm the existence of the concepts of the model in the data narrative. In Subsection 3.4.2, we further demonstrate the practicality of the conceptual model by implementing the model through a proof of concept designed to assist in the creation of data narratives. In Subsection3.4.3, we conduct an examination of the existing literature related to the practical implementation of data narrative concepts.

3.4.1 Reverse engineering data narratives

Reverse engineering data narratives consists of examining an already-existing data narrative to uncover the underlying structure and techniques used in their creation. In particular, obtaining the instance of the conceptual model of the data narratives detailed in section 3.3. These instances may be derived from the data narrative itself or from the data narrative analysis support. We consider a data narrative as a visual narrative, where a single dashboard includes various dashboard components, including textual elements and graphical figures. Similarly, we define an analysis support as a document that includes information related to the exploration, its results, the analysis goal, and the analytical questions raised during the creation of the data narrative.

In order to reverse engineer data narratives, we conduct a manual analysis of both the data narrative and its associated analysis support, with the aim of identifying the underlying concepts. Subsequently, the analysis later distilled and transformed into an algorithm, which is detailed in Appendix 1.3. This algorithm is specifically designed for identification of concepts of the model. It's worth noting that the tasks described in the algorithm can be followed for other usages as motivated before.

In this subsection, we aim to identify the concepts of the model, rather than delve into the specifics of how we conducted the reverse engineering process for the data narrative. In the following, we will detail the importance of reverse engineering of data narratives, outline the principles underlying this approach, and subsequently provide an illustrative example.

Importance of reverse engineering of data narratives. The reverse engineering of data narratives holds significant importance for several reasons:

- Understanding data narrative construction: By reverse engineering data narratives, data narrators acquire insight into the subtle art of narrating with data. They are able to understand the narrative's structure, identify concepts of the data narrative,

comprehend the analysis conducted behind the message communicated, and perceive the interconnection between messages.

- Facilitating reusability: By uncovering the underlying concepts of data narratives, reverse engineering promotes reusability. Data narrators can adapt and reuse successful narrative templates, saving time and effort while maintaining quality and consistency in reporting and communication.

- Enabling collaboration: Understanding how data narratives are constructed improves collaboration among multidisciplinary teams. Data scientists, analysts, designers, and domain experts can work cohesively, leveraging each other's expertise to create data narratives that are both informative and aesthetically appealing.

- Indexing data narratives: Reverse engineering data narratives provides valuable insights into their structural concepts, content, and organization. Leveraging these insights allows for the development of effective indexing systems, making data narratives more accessible and valuable resources for knowledge discovery and decision support.

Principles of reverse engineering. The fundamental principle of data narrative reverse engineering depends on the existence of two essential elements: a data narrative and the analysis support that goes along with it. It is impossible to perform reverse engineering without one of these documents.

We identify first the components of the visual narrative. Then, we moved to find the primary concept of the data narrative the "messages" that are being conveyed in the data narrative. Subsequently, we proceeded to identify the structural concepts inherent to the data narrative. Moving to the data narrative support analysis, we identified the factual concepts. In the final step, an alignment between the messages and related instances identified, including the findings, collectors, analytical questions and visualizations.

Example. As an example, we describe the tasks a data journalist took to craft a data narrative about the COVID pandemic in a French region. We emphasize the underlying concepts associated with each of these tasks. Tasks leading to explore the data and reflect the intention of the data journalist were deduced from the analysis support represented in the form of a notebook [41], while the tasks leading to structure and present the plot of the data narrative were deduced from the visual narrative published on Rue89 Strasbourg [42].

To ease the reading, we describe these tasks in four main parts, following the four layers of the conceptual model. Note that this does not reflect the order of steps taken by the journalist, who, in that case, started with a clear goal in mind before even collecting and exploring the data. The following order is dependent on the manual steps followed for reverse engineering data narratives.

1. *Presentational layer*: The **visual narrative** takes the form of a post that has been published on Rue89 Strasbourg [42]. This visual narrative is structured into multiple sections, each regarded as a **dashboard**. Additionally, each of these sections is subsequently broken down into what we refer to as a **dashboard component**, typically comprising textual content and often accompanied by visualizations. In the case of the dashboard component presenting the increase in mortality in March-April 2020, an interactive line chart plots, for the period and for years 2010 to 2020, the number of deaths by day and its moving average.

2. *Intentional layer*: The data journalist aimed to "analyze mortality figures in a specific area (the French Alsatian departments)", and this **goal** was clear from the start as explicitly mentioned and formulated the goal at the beginning of the notebook. Several **analytical questions** were posed that were answered through data exploration. One of the questions was: "Compare the mortality figures in 2020 with those of previous years in the Alsace departments between March 1st and April 24th, 2020". On the basis of the finding which is retrieved from the analysis, the data journalist formulated one or more **messages** that provide an answer to the analytical question. One message is: "In the Upper Rhine region, it can be said that the difference of number of deaths is undoubtedly largely related to COVID19". This message was validated by an expert (Dean of the faculty of Medicine of the University of Strasbourg). This message is composed by **characters** such as "Upper Rhine" and "COVID19", as well as **measure** like "number of deaths".

3. *Structural layer*: The data journalist addressed the general public to communicate the mortality figures in the Alsatian departments. He structured the **plot** of the data narrative into six **acts**, each composed of one or multiple **episodes**. Each episode narrates a message formulated in answering a question. One such episode narrates the message telling that COVID19 is largely responsible of the increase in mortality in March-April 2020, including the validation and an explanation by the Dean of the Faculty of Medicine.

4. *Factual layer* The data journalist collected the datasets of deceased people on the "data.gouv.fr" (open data) portal. The **exploration** of the data via several **collectors** written in Python revealed a number of key **findings**. For example, one finding is: "Between 2010 and 2019, there were an average of 1100 deaths between the months of March and April in the Upper Rhine area, while In 2020 (the year COVID19 struck) there were 2347 deaths between the months of March and April in the same area". The data journalist tried multiple visualizations, to aid the understanding of the data and retrieval of the findings.

As a result, we managed to successfully identify and delineate the key concepts embedded in the data narrative as well as the corresponding elements in the support analysis. This identification and characterization of the model in the data narrative context not only clarifies its theoretical foundations, but also highlights its applicability in practical, real-world scenarios. Essentially, this concept identification serves as a critical validation of the feasibility of the models when applied to concrete use cases.

3.4.2 A proof of concept implementing the model

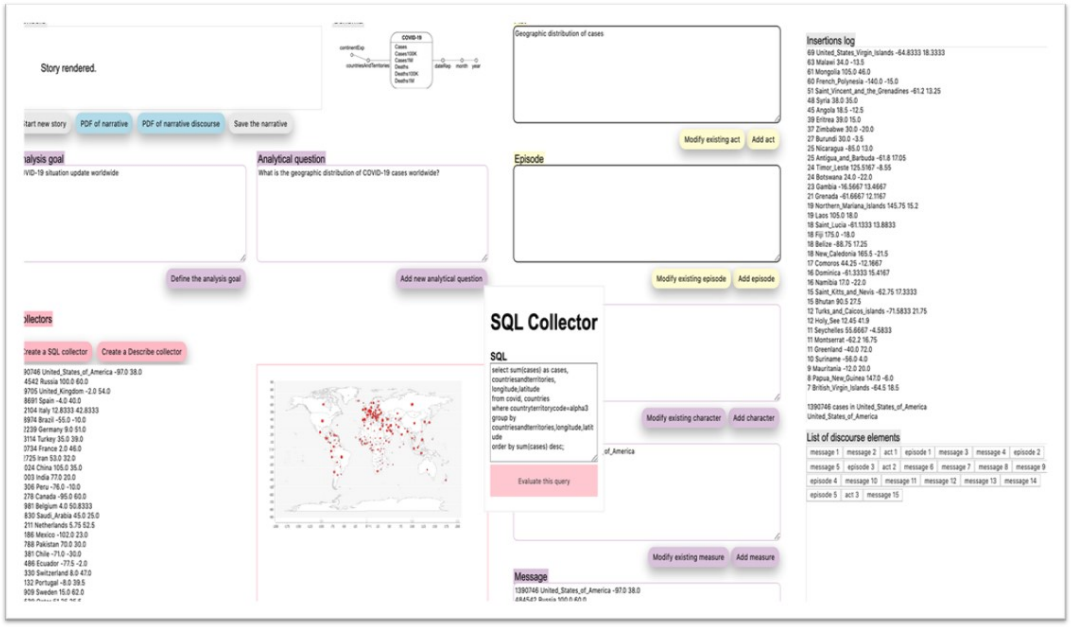


Figure 3.4: Proof of concept interface

To test the practicality of the conceptual model, we implemented a proof of concept by developing a web application that assists data narrators in crafting data narratives. This web application incorporates the key concepts of the proposed model to aid the data narrator in generating data narratives. The interface of the web application is depicted in Figure 3.4. The upcoming subsections consist of detailing the implementation of the model along with a demonstration scenario that emphasizes the application of the proof of concept for creating data narratives. The usage of the web application is detailed in Appendix 1.4.

3.4.2.1 Implementing the data narrative model

We implemented the proposed model through a simple web application to generate data narratives when exploring data. This web application is implemented in Java using Spring, d3.js, JFreeChart and Apache PDFBox.
Each concept of the proposed model is implemented as an interface. Concrete classes allow to design simple visual narrative based on (i) a factual layer that implements collectors over a relational database, and (ii) a presentational layer that renders stories as a PDF document. The user interface essentially consists of text areas where the author can declare goal, analytical questions, messages, characters, measures, episodes and acts. The application logic controls that these inputs are compliant with the model. Precisely, the author starts a new narrative with a goal, and then expresses some analytical questions. For each question, the author can try the different collectors, and inspect their answers. If the findings brought by a collector are found worth adding to the narrative, they are turned into messages, for which the author must declare at least one character and one measure. Then, an episode can be created only if it can be attached to an act

and a message that must have been declared beforehand.
The current prototype implements two types of collectors over a relational database. The first collector type allows to send plain SQL queries over the database and obtain the answer as a set of records. The second type implements the Describe operator presented in [38], which allows to enter intentional queries [163], augment the result with automatic model extraction (e.g., clustering), and render the result to appropriate charts.
As to the rendering of the narrative, the current prototype implements two types of visual narrative downloadable as a PDF file. Both use dashboard components that write the texts of acts and episodes, and the image of a chart brought by the Describe collector or produced from the result of a SQL query. Finally, all SQL collectors can be documented and returned in a SQL notebook, using the Franchise SQL notebook application[2].
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.

3.4.2.2 Demonstration Scenario

This section presents the experience to showcase the production of data narratives using interactive querying and visualizations. The demonstration is guided by a generic case study such as "As a journalist, you are investigating how COVID-19 spreads around the world." Data narrators are asked to extract relevant facts from a COVID dataset and to produce a data narrative enriched with visualizations (see Figure 3.5). The scenario mimics the data narrative published by the European Centre for Disease Prevention and Control (ECDC)[3].
We detailed each layer and component arrangement to reprint this scenario using our prototype for generating a complete data narrative. Figure 3.5 represents some screenshots of two data narrative versions: the initial one in the left part and reprinted version in the right using our prototype. The code, screenshots and a PDF of the reprinted data narrative, generated with the prototype, are available on Github[4].

Intentional layer. The data narrator starts a story by specifying the *analysis goal* of the intended data narrative: report worldwide covid-19 situation as of May 21st, 2020. This goal brings out several *characters* to play a key role in episodes narrated as "world-wide", "covid-19", "cases" and "deaths". A set of *analytical questions* is posed splitting different aspects of the goal: Which is the current covid-19 situation? How daily epidemiological curves evolve? Which is the geographic distribution of cases and deaths? These questions are answered by a set of *messages* (based on findings, see Factual layer below) such as "5 776 934 cases and 360 089 deaths were reported as of 21 May 2020". This message brings out new *characters* and *measures*, for example, "21 May 2020" and "5 776 934", which are narrated in the first episode.

Factual layer. As described in the ECDC web site, every day between 6:00 and 10:00 CET, a team of epidemiologists screens up to 500 relevant sources to collect the latest figures. The data screening is followed by ECDC's standard epidemic intelligence process for which every single data entry is validated and documented in an ECDC database,

[2] https://github.com/hvf/franchise
[3] https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
[4] https://github.com/OLAP3/pocdatastorytelling

available from the web site in XLS format. We downloaded the XLS file and inserted data in a relational table (keeping the same structure) for recreating the data exploration. The simplicity of the file structure allowed to produce all the *findings* reported by the data narrative using simple SQL queries as *simple collectors*. For instance, the daily curve of Episode 1 of Act 2 is generated with the following SQL query:

```
SELECT daterep, continentexp, sum(cases) FROM covid19 GROUP BY daterep, continentexp
ORDER BY daterep;
```

It is subsequently rendered with a bar chart. Similarly, Episodes 2 and 3 of Act 1 are produced with group by and top queries, while the last episode of Act 3 is produced by joining two group by queries. In other words, the exploration solving this narrative's analysis goal is a sequence of SQL queries over the ECDC database. These queries are available on the project Github.

Structural layer. The plot is organized in 3 *acts*, devoted respectively to narrate: a summary of the situation per continent (Act 1), daily epidemiological curves (Act 2), and geographic distribution of cases (Act 3).

Act 1 includes 3 *episodes*, narrating respectively: the worldwide summary of the pandemic, the cases reported per continent (highlighting countries reporting most cases) and the deaths reported per continent (also highlighting countries reporting more deaths).

Act 2 includes 2 episodes, narrating respectively: the daily evolution of new cases per continent, and the daily evolution of deaths per continent.

Act 3 includes 4 episodes. The first 3 narrate the geographic distribution of, respectively, cumulative number of cases, cumulative number of cases per 100 000 population, and 14-days cumulative number of cases per 100 000 population. The last episode details the number of cases, deaths and 14-days cases per country.

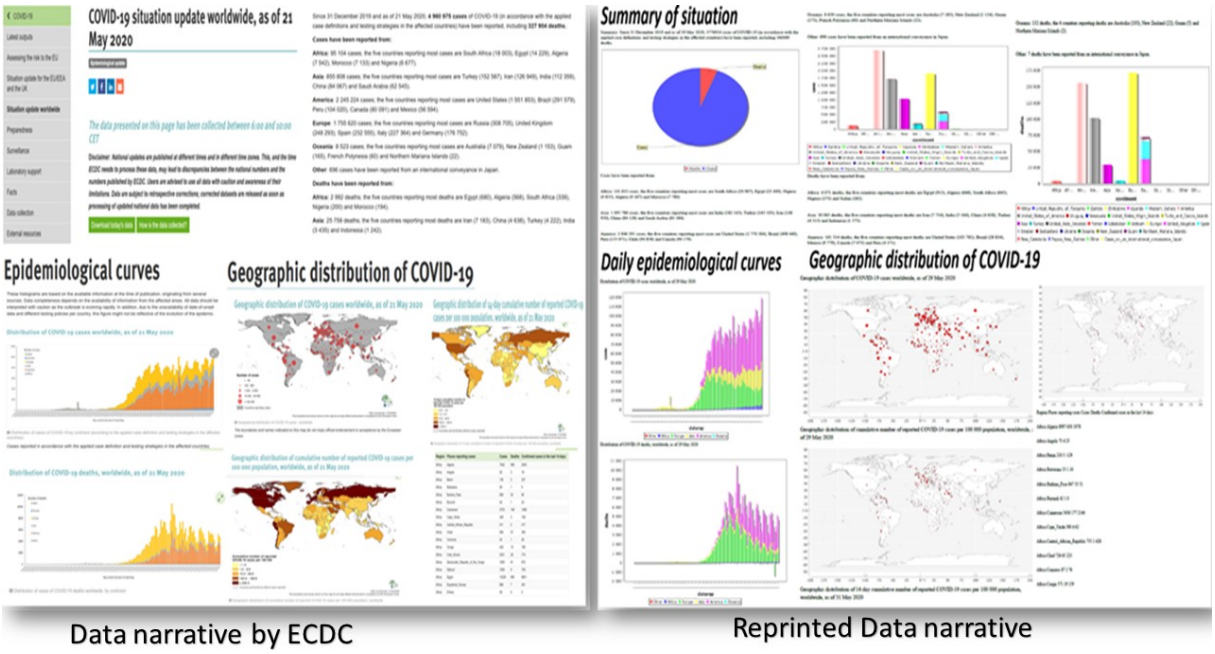


Figure 3.5: Some screenshots of two versions of covid data narrative available at https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases & https://github.com/OLAP3/pocdatastorytelling

Presentational layer. The visual narrative is published as a web page. It contains three dashboards for rendering the 3 acts of the plot. Subtitles are chosen for delimiting dashboards. Dashboard components are responsible for rendering episodes with several visual artifacts: formatted text (episodes of Act 1), bar charts with textual explanations (episodes of Act 2), maps (3 following episodes) and a table (last episode).

3.4.3 Towards automating data narration

We offer a detailed examination in Appendix 1.2, which stems from a thorough exploration of the current body of literature pertaining to the practical application of data narrative principles. Our aim was to gain a thorough understanding of how these concepts can be implemented in real-world scenarios, as well as to identify any challenges or limitations that have been encountered. By summarizing the insights gleaned from this review, our analysis revealed an abundance of research devoted to implementing the factual and presentational levels of data narratives. Notably, we observed that the communities of exploratory data analysis and data visualization have made significant contributions to the real-world implementation of these concepts.
To automate the intentional layer of data narratives, intentional operators [161, 163] and narrative patterns [9] can be employed to facilitate the automation of the data narrator's analytical questions and messaging. However, more work is needed to effectively capture the subjectivity of the data narrator, including their thought and analysis goals. Our review of the literature revealed a dearth of research in this area, particularly with regard to the implementation of message characterization in terms of characters and measures.
In a larger sense, the organization of a sequence of messages addressed in the structural layer is handled globally by the graph driven approach [80], storytelling timelines [24] and automatic tools [68, 149, 150, 168] that generate a data narrative without providing deep details on the plot sub parts in terms of acts and episodes. To achieve this, the plot sub parts can adhere to the film's traditional structure in terms of acts and episodes, as described in [68] or the scene notion described in [149].

3.5 Conclusion

This chapter introduced a conceptual model for data narratives, by extending a classical model of narrative [32] to reflect the transition from raw data to the visual rendering of messages derived from data analysis. Our model translates the fundamental concepts of narrative to their respective counterparts when it comes to data narratives and involves the concepts concerning the collection of data, extraction of key findings and the corresponding messages to the audience, structuring of a presentation of these findings and the ultimate presentation via visual -or other- means via a set of dashboards. To showcase the model, we implemented a proof of concept as a web application to assist a data narrator during crafting a data narrative, while interactively exploring a database.
The model for data narratives is considered as the foundation of the data narrative in terms of the definition of the data narrative [65, 133] and the representation of the data narrative [49, 94, 94, 120].
The effectiveness of our model extends beyond data narratives and has been successfully applied in various domains, including query language and model-driven engineering.

For example, Wang et al. [167] developed a versatile system that generates natural language descriptions of query execution plans, aiding in understanding the steps involved in executing a query. Our model's layers play specific roles in this context: the factual layer represents the query execution plans using language-annotated operator trees, enabling manipulation for generating data narratives; the intentional layer focuses on the content of the story by describing various operators to fulfill the goal of comprehension for learners; the structural layer organizes the narrative's plot by arranging the steps in an audience-friendly manner; and the presentation layer handles the visual presentation of the narrative.
In the field of model-driven engineering, Calegari [27] expanded upon our model to propose a narrative metamodel, which offers abstract models for data narratives. The layers of our model were instrumental in the model-to-text transformation to HTML and Jupiter notebooks, providing a structured and informative representation.

Chapter 4

A process for crafting data narratives

Tony Gaskins said: '*Trust the process. Your time is coming. Just do the work and the results will handle themselves*'. In alignment with Gaskins' quote, the objective of this chapter is to introduce a well-defined process for crafting data narratives.

This process is carefully designed by drawing insights from an extensive analysis of scientific literature and leveraging the practices employed by expert data journalists. By identifying and delineating the crucial phases and workflow involved in the creation of a data narrative, this process aims to serve as a guideline for data narrators, ensuring a structured approach to crafting compelling data narratives.

Just as Tony Gaskins encourages embracing, trusting, and enjoying the process, this chapter seeks to inspire data narrators to embrace the process of data narrative creation, trust in its activities, and find joy in the journey of crafting engaging narratives supported by data.

Contents

4.1 Introduction

Despite the diversity of activities, sometimes even conducted by different people with varied professions and skills, as far as we are aware, there is no workflow or tool for supporting the crafting of data narratives.
To fill this gap, we reviewed the literature around the process of crafting data narratives, as documented in Chapter 2 and specifically in Section 2.3. On top of that, we conducted a survey with data journalists in order to understand how they craft a data narrative. As an outcome of the former and as discussed in Section 2.5, we found that the research communities globally agrees in the fact that the crafting process includes three main phases:

- The *analyzing* phase that handles the activities of exploring data and retrieving findings,
- The *structuring* phase that includes the activities to organize the plot of the narrative in an understandable way and,
- The *presenting* phase that covers the activities to convey visually the plot of the data narrative.

At the same time, our bibliographical study revealed the absence of a comprehensive and well-founded process that covers the main activities of data narrative crafting, specially those dealing with user intentions and their tight relation to data analysis.
Apart from the bibliographical study, the conducted survey allowed us to observe the crafting workflows regularly followed by 18 data journalists, and we contrasted them to the literature. It turned out that journalists follow the same three phases, mostly in a linear way, attaching less attention to the structuring phase, while spending more time in the analyzing phase. Data journalists, on the other hand, invest their time in intentional activities when crafting data narratives.

These considerations from the literature study and the survey with data journalists enabled us to identify the activities (and their chaining) for crafting data narratives. Based on those, we propose a comprehensive and well-founded process that (i) covers the whole cycle of data narrative crafting, from exploration of the 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, detailed in Chapter 3, that clarifies the concepts involved in the process

The scope of our process targets the population of data journalists or any other data enthusiast that craft data narratives out of existing data. The reason for proposing the process is exactly the observed discrepancy between literature and practice, with omissions of important parts from both sides. *Thus one significant contribution of our work is the explicit treatment of all the steps that should be involved in the process, as well as providing detailed descriptions of the activities that accurately reflect the intention of the data narrator.* Secondly, apart from providing a methodological guidance, *our process*

can enable the support of the process via tooling. Indeed, there is a lack of integrated tools covering the whole crafting process and recommending actions to less-experienced 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 a conceptual model and a process that are both consensual.

In [130], the process's origin, motivation, and initial presentation are described. Additionally, a significant portion of the material in this chapter was previously published in the Information System Frontiers journal [54].

The chapter is organized as follows: Section 4.2 discusses the survey we conducted with data journalists. The proposed process is described in Section 4.3. Then we focus on detailing the question answering phase of the process in Section 4.4, primarily because there is a notable lack of literature devoted to this particular phase. Furthermore, in their practices, data journalists pay close attention to this phase. Section 4.5 describes our experiments and Section 4.6 concludes.

4.2 Data journalist practices

We report the results of a survey [30] (in French), aiming at investigating the professional practices of data journalists.

Journalists	1	2	3	4	5	6	7	8	9	10	11	12
1	Choose datasets	Query data	Implement visualization									
2	Choose datasets	Analyze data	Write article									
3	Analyze data	Validation	Write and shape the article									
4	Analyze data	Sort messages	Choose focus	Shape the article								
5	Analyze data	Choose visualization	Implement visualization	Test visualization								
6	Formulate hypothesis	Choose datasets	Verify hypothesis (Analysis)	Write and shape the article								
7	Formulate hypothesis	Collect data	Query data	Draw patterns	Interviews (validation)							
8	Formulate question	Choose datasets	Analyze data	Interviews (validation)	Write article							
9	Overview of data	Choose subject	Prepare data	Analyze data	Implement visualization	Write article						
10	Choose subject	Formulate hypothesis	Verify hypothesis (Analysis)	Design visualization	Implement visualization	Write article						
11	Idea of subject	Choose datasets	Prepare data	Analyze data	Choose visualization	Write article						
12	Choose datasets	Prepare data	Query data	Draw findings	Choose visualization	Write article	Choose final visualization					
13	Formulate hypothesis	Choose datasets	Collect data	Prepare data	Analyze data	Validation (cross-checking)	Implement visualization					
14	Choose subject	Choose dataset	Collect data	Analyze data	Refine subject	Discussion (Check finding)	Search further datasets	Analyze data	Interviews (validation)	Prepare visualization	Choose form of article (paper/web)	Write article

Figure 4.1: Sequence of activities reported by journalists

The survey consisted of 32 questions[1] (in French). Note that for some questions more than one answer was possible, and that journalists could leave the questions unanswered. The survey was answered by 18 data journalists from 14 French regions, who

[1] https://tinyurl.com/ynjzjs63

have worked on a big variety of topics, including elections, environment, cinema, terrorism, paradise-papers, real estate. For nearly 50% of them, data narration is at the core of their professional activity, and is occasional or marginal for the others. Concerning training, 56.3% studied social sciences, 18.8% studied sciences and 24.9% graduated from law or journalist schools. One of the journalists works for the International Consortium of Investigative Journalists (ICIJ), 5 of them work for the national press, and the 12 remaining work for the regional press. Regarding their general working habits, 75% of them work alone. They usually work on open data (72.2%) and more specifically on data from public institutions (44.4%). They consume from minutes to months during the data narration and use different tools during data exploration, such as spreadsheets (93.8%), scripts (50%), notebooks (18.8%), PowerBI-like tools (31.3%) and some machine learning tools (28.6%).

Two main questions were asked on their data narration practices.
For the first one, "How does a data story's subject emerge?", multiple answers were possible. The answers showed that the goal, or subject, of an article emerges from: an idea to be confirmed by data (68%), a dataset which needs exploration to reveal important facts (68%), a refinement of the subject while exploring the dataset (48%).

The second, open question was: "What is the general workflow you apply for data narrative crafting?". Fig. 4.1 sketches the answers provided by 14 of the 18 journalists, where activity names summarize journalists' descriptions of their main activities[2], rows correspond to journalists and column numbers reflect the sequence of activities.
We color these activities according to the layers of the conceptual model: factual (pink), intentional (purple), structural (yellow) and presentational (blue). Gray-colored cells indicate that the activity may overlap structural and (more probably) presentational tasks. In addition, activities concerning the checking of findings and the validation of messages (namely interviews, validation or cross-checking), aiming at transforming a factual object into an intentional one, are in between the factual and intentional layer. Similarly, visualizations are used both in the factual layer, to understand data and retrieve findings, and in the presentational layer, to choose the most suitable one for communicating findings to the audience in a visual manner.

[2]Since the question was open, we homogenized the answers and grouped them into few categories.

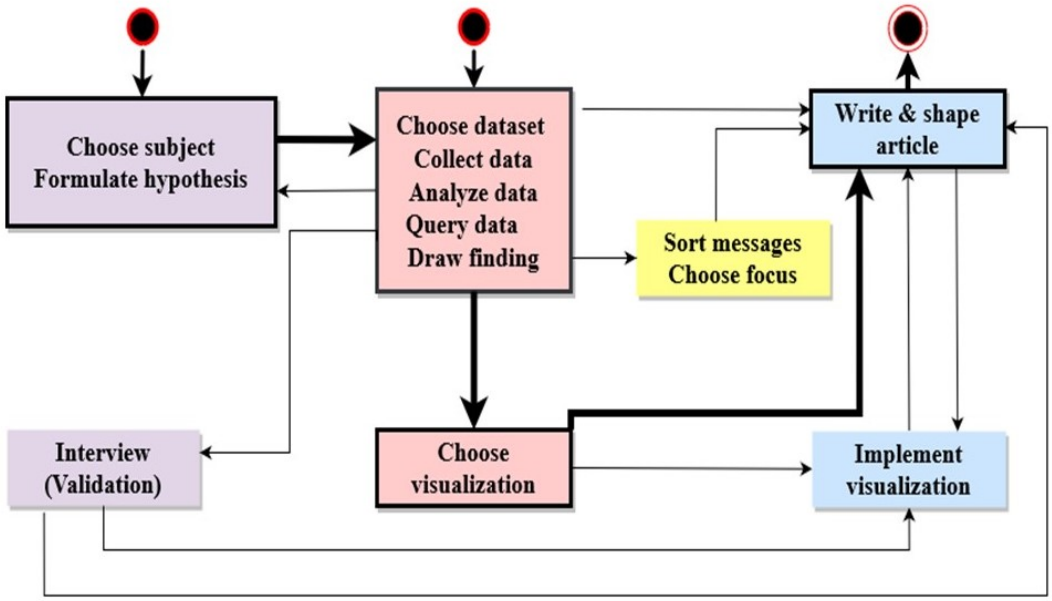


Figure 4.2: Activity diagram for activities reported by data journalists

We have abstracted these sequences in the form of an activity diagram depicted in Figure 4.2. Most frequent paths are highlighted by larger arrows.

Lessons learned. Fig. 4.1 shows that many activities under different names aim towards the same action, and that different paths can be followed by journalists when crafting a data narrative. *The figure also shows a preponderance of activities from the factual and the intentional layer.* The activity diagram depicted in Figure 4.2 shows that journalists enter the workflow either in the factual layer, i.e., by exploring a dataset, or by the intentional layer, i.e., having at least a vague idea of the subject. After this, the workflow becomes mostly linear, with some movements between the factual and intentional layers. Usually, data journalists start writing their articles once the analysis phase is over, and there is no backtrack once the presentational layer is entered.

Notably, the journalists attached less importance to the activities of the structural layer. 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. For those that mention it, the activity "write article" includes several hidden details concerning the organization of messages that should be communicated, the visual presentation and communication of the analysis results.

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 the presentational part. On the other hand, when it comes to the literature (see Chapter 2), 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 the creation narratives that are more complete and intuitive.

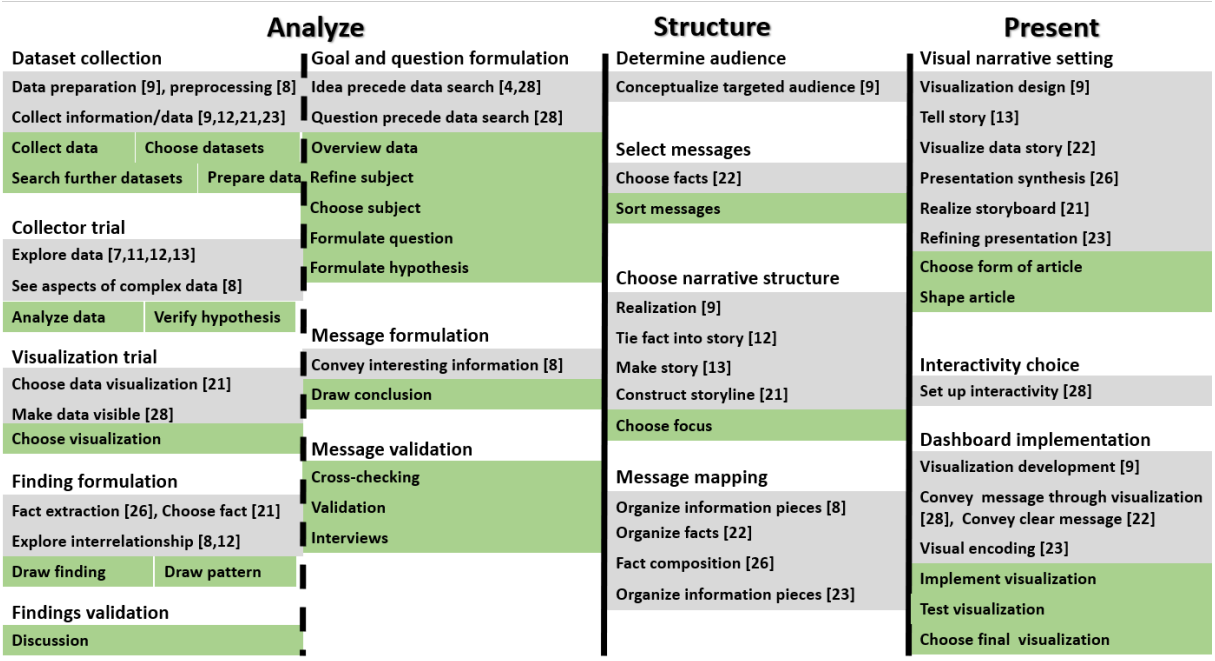


Figure 4.3: The main activities for crafting data narratives identified from the literature (in gray boxes) and a survey with data journalists (in green boxes)

4.3 A process for crafting data narratives

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

4.3.1 Requirements

First of all, we note the absence in the literature of a whole workflow for crafting data narratives, including all the activities identified from the literature (see Section 2.3) and data journalist practices (see Section 4.2).

Figure 4.3 depicts the activities as phrased in the literature (in gray boxes) and by journalists (in green boxes). We group those referring to the same task and propose new names (the titles were displayed in a larger font size in Figure 4.3) which are consistent with the conceptual model of Figure 3.3.

In more details, most authors [35,47,92,98,149,150,168,169] agree that data narration process includes three main phases: (i) *analyze*, (ii) *structure*, and (iii) *present*.

The survey reveals that the data journalists agreed with the literature, especially on the phases (i) and (iii). In Figure 4.3, activities are grouped according to these phases. We remark that activities pertaining to the factual and intentional layers of the conceptual model are mixed in phase (i). In addition, while the literature rarely mentions the activities pertaining to the intentional layer, these activities are often pointed by data journalists. Furthermore, as we explained in Chapter 3, the substance of a story, representing the narrator's intention in reporting the story, is a constituent of the data narrative [32].

Conversely, while the journalists did not attach much importance to the activities of the

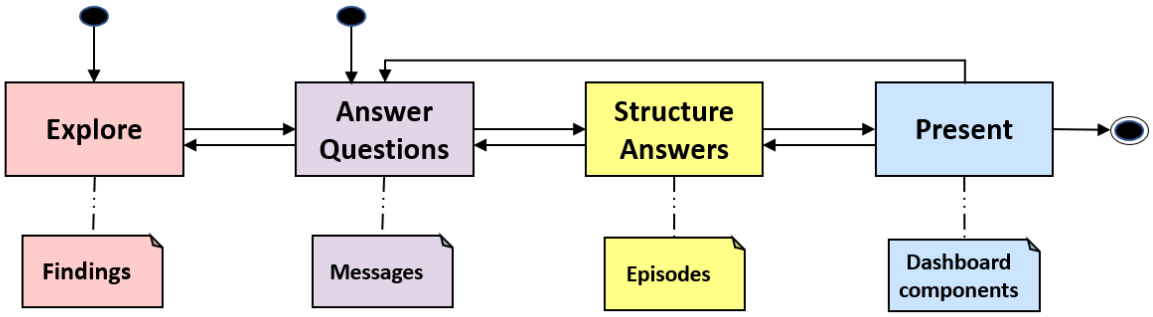


Figure 4.4: The process of data narrative crafting

structure phase, this aspect is emphasized in the literature. Finally, as noted in [98, 169], the narrator should have the possibility to move freely back and forth between the different phases of data narration. However, this movement should not prevent that different groups of activities could be conducted by different persons with different profiles. These groups of activities, identified by layers in the conceptual model of Chapter 3, should be as isolated as possible.

To summarize, a comprehensive workflow for crafting data narratives 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, which are depicted in Figure 4.1,
- (R_2) cover the activities of the three phases identified from the literature (see Section 2.3.1 of Chapter 2),
- (R_3) allow the free back-and-forth transition between phases,
- (R_4) clearly delineate the different layers of the conceptual model introduced in Chapter 3 within its activities.

4.3.2 The process of crafting data narratives

In this subsection, we propose a comprehensive process for the crafting of data narratives that covers the activities and paths proposed in the literature and reported by journalists (requirements R_1 and R_2), while also being founded upon and coherent with the conceptual model (R_4) and allowing the back and forth movement between its phases (R_3).

The phases of the process are showed in Figure 4.4. All phases are accompanied by the resulting outcomes, which are exactly the basic constituents of our conceptual model (R_4). Note that, the incomes of the *structure* and *present* phases are more than just the basic constituents; rather, they are the organization of episodes and dashboard components (see subsection 2.1.1 in Chapter 2). We retain the same coloring (pink for factual exploration, purple for intentional question-answering, yellow for the structuring of the answers of the intentional questions into a plot, and blue for presentation). Observe 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 Figure 4.1, the process flexibly starts either with the existence of a data set to be explored for findings, 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 specificities of the task on hand, alternating the exploration of data, answering questions by deriving messages, structuring the answers and presenting visually the structured answers.

In any case, the identification of the answer 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.

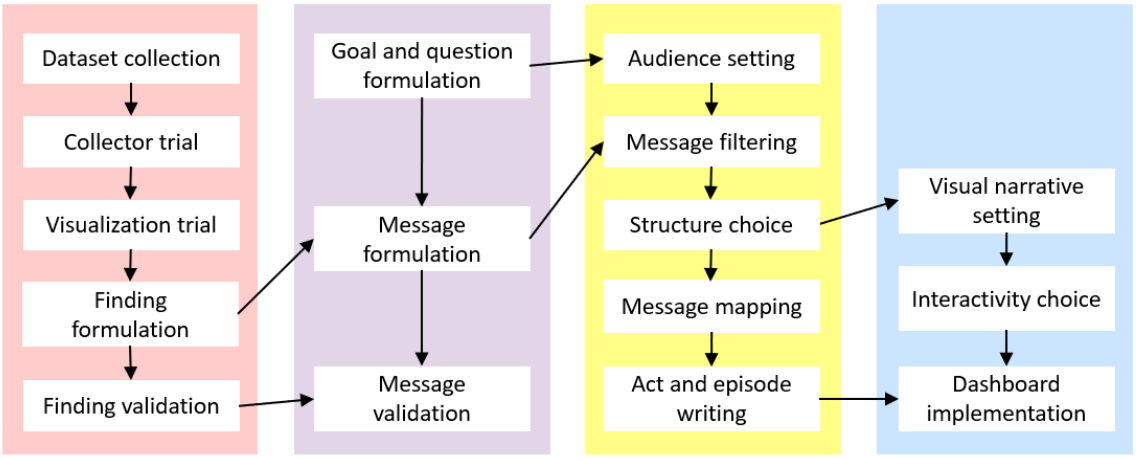


Figure 4.5: Activities for data narrative crafting (→ indicates a " depends on" relationship)

The following paragraphs present the activities associated with each phase. These activities are abstracted from the literature and survey results (shown in Figure 4.3). A new activity *act and episode writing* is added in order to explicitly state the task of conceiving, naming and eventually writing some notes about 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 for data narratives, which are implicit both in the survey and the literature.

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 Figure 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.

The next paragraphs detail the phase of data narrative crafting and outline the activities associated with each phase.

Explore The exploration phase, handling the factual layer, includes 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, which is typically done via statistical tests, but also by discussing and crosschecking with additional data sources (as done by data journalists) and confronting with the state of the art (as done by data scientists [110]). Note that some findings may lead to additional analysis, triggering more collectors and visualisations, or even the collection of more datasets. The exploration phase is time-consuming (data journalists measure it in days or even in months).

Answer questions This phase, neglected in the literature, handles the intentional layer and includes activities for (i) formulating goals and questions, (ii) drawing messages from findings, and (iii) validating messages. It supports explicitly the data narrator 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 Figure 4.3), we propose a message formulation activity, concerning the derivation of messages from findings, and the identification of characters and measures (the relevant constituents of messages (see subsection 2.1.1 in Chapter 2) to be highlighted to the audience.

Structure answers The structure answers phase, often underestimated in in the data journalists practices, handles the structural layer and includes activities for organizing the plot of the narrative in terms of acts and episodes. Plot setting starts by (i) determining the audience, (ii) possibly selecting a subset of messages for such audience, and (iii) choosing an appropriate narrative structure. Then, (iv) messages are mapped to acts and episodes. In more details, these activities allow the arrangement of the messages into different layers: an act which is a major piece of information, and is composed of several episodes that are of lesser importance on their own [128]. Remember that the result of the structuring is an *episode*, which is the annotation of a message (which has a simple structure and labeling) 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 Figure 4.5, the existence of a specific activity to make the actions of writing acts and episodes explicit. Such activities 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 (ii) 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.

4.3.3 Scenarios for crafting data narratives

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 common unfolding scenarios described by practitioners or observed. Scenarios are identified based on the following: the study about data journalist practices described in Section 4.2, the analysis of several data narratives and their associated processes published by data journalists, and the observation of several practitioners (as will be detailed in Section 4.5). The scenarios are sketched in Figure 4.6 by means of regular expressions.

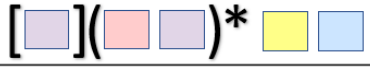
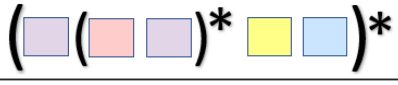
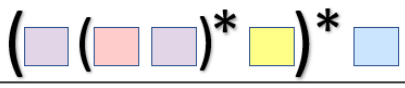
Scenarios	Regular expressions for crafting data narratives
Exploratory	
Pre-canned	
Question-by-question	
Delegated-presentation	

Figure 4.6: Regular expressions representing the unfolding of phases in different scenarios for crafting data narratives. Colored boxes represent phases, respectively pink for Explore, purple for Answer questions, yellow for Structure answers and blue for Present. * denotes repetition, ( ) is used for grouping, and [ ] indicates an optional element.

An *exploratory* scenario is commonly observed when the analyst 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 [41] describing the process followed to build a data narrative [42] about covid pandemic in a French region. 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[3]. 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.

A *question-by-question* scenario consists in chaining all phases one question at a time.

[3]https://www.fastly.com/blog/election-2020-a-data-story-in-three-parts

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) among the previous phases, preparing the plot. This scenario was reported by several interviewed data journalists [30].

4.4 A focus on the answer questions phase

The *Answer questions* phase, which has previously been neglected in the literature, deals with the intentional layer and how it relates to the factual, structural and presentational layers. In this section, we detail the *workflow* for the answer questions phase that 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 such as data scientists, data analysts, etc.

The workflow is modeled by three activity diagrams, respectively in Subsections 4.4.1, 4.4.2, and 4.4.3. For readability purpose, each diagram highlights one of the 3 activities of this phase (goal and question formulation, message formulation and message validation), by detailing the incoming and outgoing arcs to facilitate the understanding of the succession of activities. Since we focused on showing the incoming and outgoing connections of a single activity, some activities might appear disconnected because they lack direct connections to the activity that is currently being discussed. The paths from the activity are depicted with bold arcs, while the paths to the activity are depicted with regular arcs.

4.4.1 Goal and question formulation

The *Goal and question formulation* activity reflects the high-level intentions of the data narrator, and therefore helps identifying potential data for exploration and influences the structure of the data narrative. Figure 4.7 depicts its flow.

Incoming arcs. This activity can be the first of the process (top incoming arc), when a goal, and possibly some questions, are initially formulated.

Conversely, it can be entered after some data exploration (incoming arcs from the left). Indeed, while exploring and visualizing a collector's output in order to gain a deeper understanding of the data, possibly, new analytical questions can be posed, seeking for different information and inviting for further exploration. The same may arise after a finding is formulated and validated. Furthermore, the selection of a previously asked question that was not (completely) answered is also possible.

Later, when data exploration has lead to the formulation and validation of messages, new questions may appear or old questions may be reformulated, so this activity can

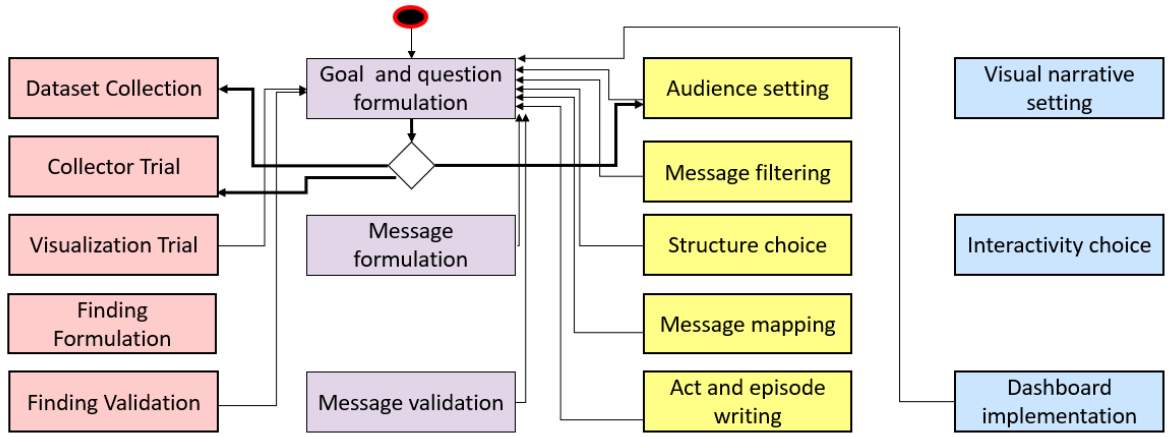


Figure 4.7: Goal and question formulation flow

also be reached after message formulation or message validation (bottom incoming arcs). Many iterations can follow in this way, specially in exploratory and question-by-question scenarios, allowing the expression of new questions after some exploration and some messages, which in turn, will trigger more exploration.

Finally, after partially structuring the plot or even implementing visualizations, the data narrator can come back to formulate new questions (incoming arcs from the right). For example, while mapping messages to acts or writing acts and episodes, a new question can be asked if some missing aspects are detected, in order to complete the plot of the data narrative. In a question-by-question scenario, this come back to the intentional phase, looking for the following question to deal with, is particularly frequent.

Outgoing arcs. After formulating goal and questions, the natural sequel is to explore data, either by collecting datasets (especially the first time) or just trying or reusing collectors (outgoing arcs to the left). If formulating goal and questions was the first activity in the crafting process, dataset collection is necessary to start exploration. In addition, even after some exploration, a question may require the collection of additional datasets if the available ones lack some information. Conversely, this activity may directly be followed by collector trial to directly explore existing datasets.

Once a goal and some questions have been formulated (and typically some messages have been validated), the data narrator can set the appropriate audience for the data narrative (outgoing arc to the right) and possibly start structuring.

4.4.2 Message formulation

The *Message formulation* activity concerns the derivation of messages from findings, intended to answer analytical questions. Its flows are depicted in Figure 4.8.

Incoming arcs. Message formulation naturally takes place after finding validation. Indeed, during data exploration, some findings arise, which are in turn validated via cross-checking or statistical testing. These findings are then compiled into messages to the audience, highlighting characters and measures. This is the unique incoming arc to this activity.

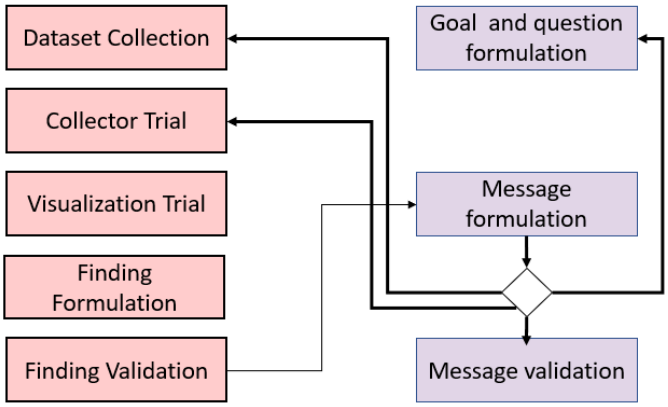


Figure 4.8: Message formulation flow

Outgoing arcs. After formulating a message, there are many options to continue the data narrative crafting. Message validation (outgoing arc to the bottom) is the more natural, allowing for a direct verification of the validity of the message. But note that the validation of messages is not required to be done immediately after they are formulated. Indeed, some narrators prefer to validate all messages together, especially when such validation involves external experts. This can help save time, allowing the data narrator to explore more data before sending a set of messages to be validated.

The data narrator may prefer to continue exploring data, in order to find additional substance to answer the analytical question at hand. This can be done by collecting a new dataset or trying a collector to further explore an existing dataset (outgoing arcs to the left).

In turn, the narrator can choose to express a new question, or select another existing question to treat (outgoing arc to the top). This latter flow was already explained in previous subsection.

4.4.3 Message validation

The *Message validation* activity ensures the validity of messages, typically asking for expert's advice or comparing to the state of the art. Figure 4.9 illustrates its flows.

Incoming arcs. Message validation comes after message formulation (unique incoming arc). As explained in the previous subsection, this can occur either one by one, immediately after formulating each message, or all messages at a time, after formulating several messages.

Outgoing arcs. After validating a message, there are many options to continue the data narrative crafting. A data narrator may, for instance, pose a new analytical question or continue answering the same question by collecting new datasets or applying a collector to explore an existing dataset.

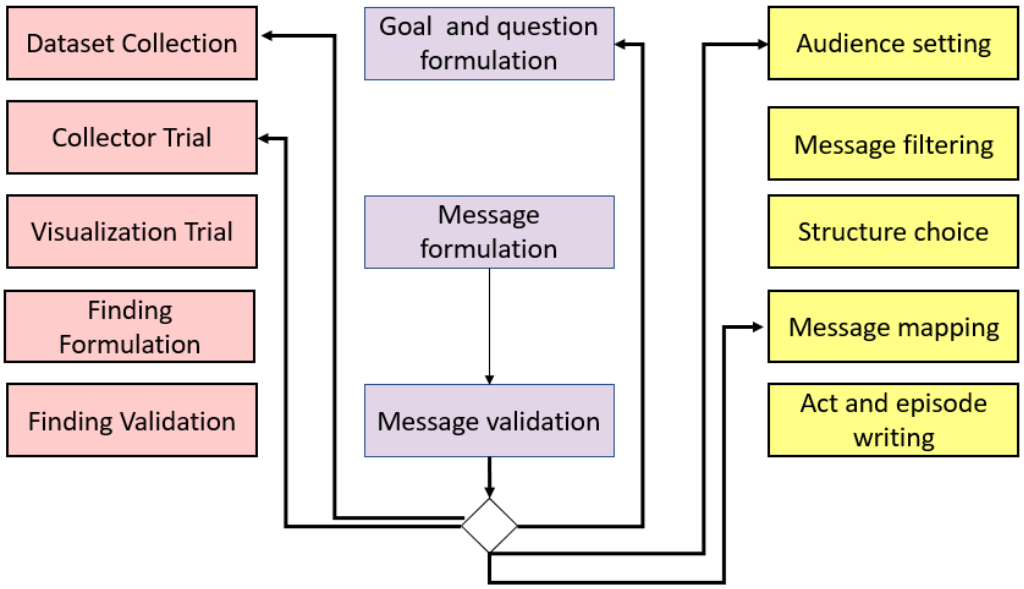


Figure 4.9: Message validation flow

Also, the data narrator can pass to the *Structure Answers* phase, by setting the audience of the data narrative, or mapping the validated message to acts. The former is done in the first passing to the structure answers phase, typically when enough messages are validated and the analytical questions are reasonably answered, having a good idea of the story to be told. The latter is done when the plot is already initiated and some additional messages arrive, in order to map them to acts and start writing.

4.5 Experiments

To validate experimentally the proposed process, we organized two challenges and analyzed several publications describing crafting processes followed by data journalists and data scientists.

The challenges aim at answering two questions: (i) Does the process *cover* all necessary activities performed by data narrators? and (ii) To what extent do the process *phases contribute* to the quality of the data narratives?

The analysis of published processes aim at answering the question: (iii) Is the proposed process consistent with the reported ones?

Subsection 4.5.1 addresses 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. The second question is addressed in Subsection 4.5.2. An experienced data journalist assessed the quality of data narratives crafted by Master students, and judged the completion of each process phase. Concretely, we investigate the correlation among phase completion and narrative quality.

Subsection 4.5.3 describes our analysis of some published narratives and the associ-

Teams	A	B	C
Topics	Migration	Vegetation	Woman-named streets
Visual narratives form	Video	Notebook	Interactive book
Starting from	Vague idea	Precise idea	Vague idea

Table 4.1: Characteristics of the crafted narratives observed during the workshop

ated processes as documented by the narrators. Concretely, we investigate whether the proposed process is coherent with the documented ones, highlighting the scenarios that better represent them. Subsection 4.5.4 details a survey conducted that highlights the importance of having a process behind any data narrative.

4.5.1 Coverage

We organized a one day challenge called "Narrating Rennes by the data"[4] with data enthusiasts and data scientists, aiming at producing data narratives using the open data of the French city of Rennes[5]. We outline the methodology taken as well as the results of this challenge. **Methodology.** 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, the observers 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 proposed in this paper; only the observers were aware of it. The details of our analysis and the narratives produced (in French) are publicly available[6]. A prize was awarded to the best one.

Table 4.1 reports, for each team, the topic of the data narrative, the form of visual narrative, and its starting point.

Figure 4.10 lists, for each team, the sequence of activities observed during the workshop, which are also sketched as a sequence of boxes, colored as the phases of our process.

Results. The main observation one can make from Figure 4.10 is that the proposed process covers the data narrators activities and their chaining, whatever their initial idea, the topic chosen, or the style of visual narrative. In more details, we found that each group 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 activity arrived a bit later after the trial of several collectors). This did not prevent the groups to continue with the narrative crafting, and more importantly, the observers noted that no activity conducted by the group was absent from those listed in Figure 4.5.

Interestingly, we remarked that two teams started with a vague idea of the topic they

[4]Sponsored by CNRS https://www.madics.fr/event/titre1617704707-3351/#madona.

[5]https://data.rennesmetropole.fr/

[6]https://drive.google.com/drive/folders/1zDzP_ndSlQUJCbtFMVzJDnIbyXK1D2_l?usp=sharing (in French)

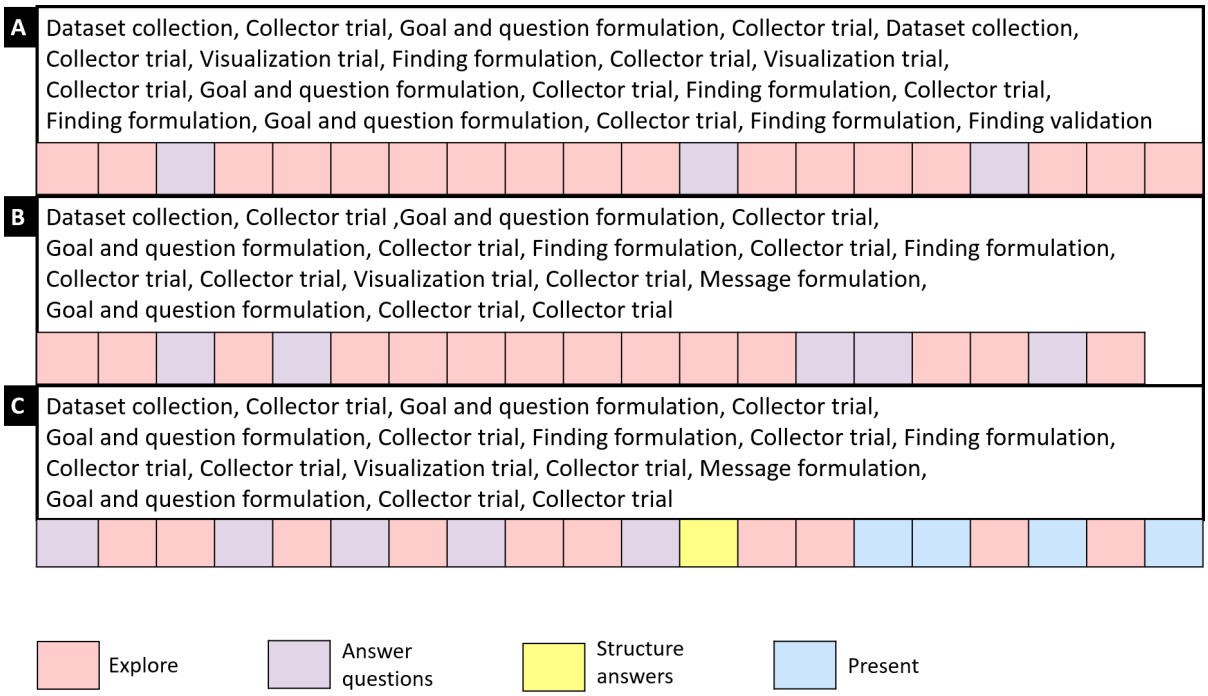


Figure 4.10: Activities for crafting data narratives observed during the workshop

wanted to treat, which was refined after many iterations among data collection, data analysis and question formulation. This clearly correspond to an exploratory scenario. Furthermore, we identified 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 of them used a unique timeline for structuring their narratives, which were rendered with varied styles.

We can also note that our proposed process remains tailored for the task at hand. Indeed, the activities reported in Figure 4.10 cover almost all the activities of our process. Activities that were not reported in Figure 4.10, particularly those related to structure answers and present phases, were likely completed after the workshop.

4.5.2 Phases contribution to narrative quality

For assessing the relationship between process phases and narrative quality, we asked an experienced data journalist to evaluate a set of data narratives, assessing both their quality and the perceived phase completion. Below, we present the methodology conducted, along with the results obtained in this study.

Methodology. Narrative quality was assessed on a scale from 1 (lowest) to 7 (highest), using 3 criteria (previously proposed in [52]): (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), was 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 had been completed, based on how well the data narratives translate the expression of the intention of the data narrator and

how much of the subject was investigated. In the same way, the data journalist assessed how much of the *structure answers* and *present* phases had been completed. The *Explore* phase was omitted from the evaluation because data narrators submitted only the data narratives without providing any documentation for the exploration they had conducted while crafting the narratives.

To this end, we implemented a challenge for Master students in Computer Science, specialized in data analysis. 44 students participated in the challenge, 14 of the first year of the 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. Students were asked to craft a data narrative about fatal encounters in the USA, using an open dataset[7]. 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.

		assessed quality				perceived completion		
		Info	**Comp**	**Expe**	**Avg_Q**	**C_ans**	**C_str**	**C_pre**
All	**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
M1	**Min**	1	1	1	1.00	1	1	1
	Max	5	6	5	5.33	5	5	7
	Avg	3.00	3.43	2.86	3.10	2.57	3.29	4.00
	Stddev	1.41	2.07	1.35	1.58	1.27	1.50	2.00
M2	**Min**	2	1	1	1.89	1	2	2
	Max	5	6	5	5.33	6	6	7
	Avg	3.53	3.71	3.35	3.57	3.18	3.82	4.24
	Stddev	1.01	1.26	1.11	0.92	1.51	1.33	1.35

Table 4.2: 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 criteria.

Results. The results of the evaluation are reported in Table 4.2. Both M1 and M2 students produced narratives graded from 1 to 6, with similar average quality (with less deviation for M2 students), despite their background differences. Students were observed during crafting, and some of them, especially the M1, were asked to indicate the sequence in which they completed the activities depicted in Figure 4.5. This helped them to start, particularly having to write down the analytical questions that guided the data analysis, and to write down messages and initially consider how to structure them.

As to the different phases, the present phase was better completed than the two others. In addition, we measured the correlation (using Pearson correlation coefficient) between

[7]https://fatalencounters.org/

the average quality (Avg-Q in Table 4.2) and the completion of the three phases. The correlations were, 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 was 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 influences narrative quality at least as much as the other phases, which confirms our claim about its importance for data narrative crafting.

4.5.3 Comparison to documented processes

In this subsection, we compare our process to other documented processes and present the methodology conducted, along with the results obtained in this study.

Methodology. Actually, we study four works that documented (at least some portions) of the crafting process followed to produce a data narrative, namely, (i) a data narrative [8] about *the climate crisis in the Sahel* documented by a data journalist in the form of a blog [9], (ii) a data narrative [10] about *tennis betting data* documented by an investigative data reporter for BuzzFeed News in the form of a sport news[11], (iii) a data narrative [42] about *the COVID pandemic in a French region* documented by a data journalist in the form of a notebook [41], and (iv) a data narrative about the *tuberculosis pandemic in Gabon*[12] documented by a data scientist in the form of a research article [110].

Some of the works merely described the main activities accomplished, without detailing every iteration adopted during the crafting process. Other ones only detailed the early phases of the process.

For three of the narratives, namely those about Climate crisis in Sahel, Tennis betting data and Tuberculosis pandemic, the process was clearly described. For analysing them, we just needed to match the activities listed by data narrators to those of our process, highlighting the flow of activities.

However, for the narrative about the Covid pandemic, the process is reported in a Python notebook, mostly detailing the 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 manual instructions applied during the reverse engineering of the data narrative are represented in the form of an algorithm for manual concept identification, which is detailed in Appendix 1.3.

We recall the steps for reverse-engineering the data narrative: we began by identifying the goal and analytical questions, which were explicitly stated at the beginning of the notebook. Collectors were implemented as python code, the results of which were commented. We identified findings within the data journalist's comments. Then we attempted to locate these findings within the data narrative, looking both for text explaining the finding

[8] https://data.humdata.org/visualization/climate-crisis-sahel/
[9] https://tinyurl.com/ynjzjs63
[10] https://tinyurl.com/zxwf34xt
[11] https://tinyurl.com/wa4jaenj
[12] https://www.youtube.com/watch?v=u_KoBWc_qJU (in french)

and a visualization similar to the collector output. When we succeeded in matching some textual or visual artifacts, we took them as the formulation of a message. The activities of structuring and presenting were not mentioned explicitly by the data journalist.

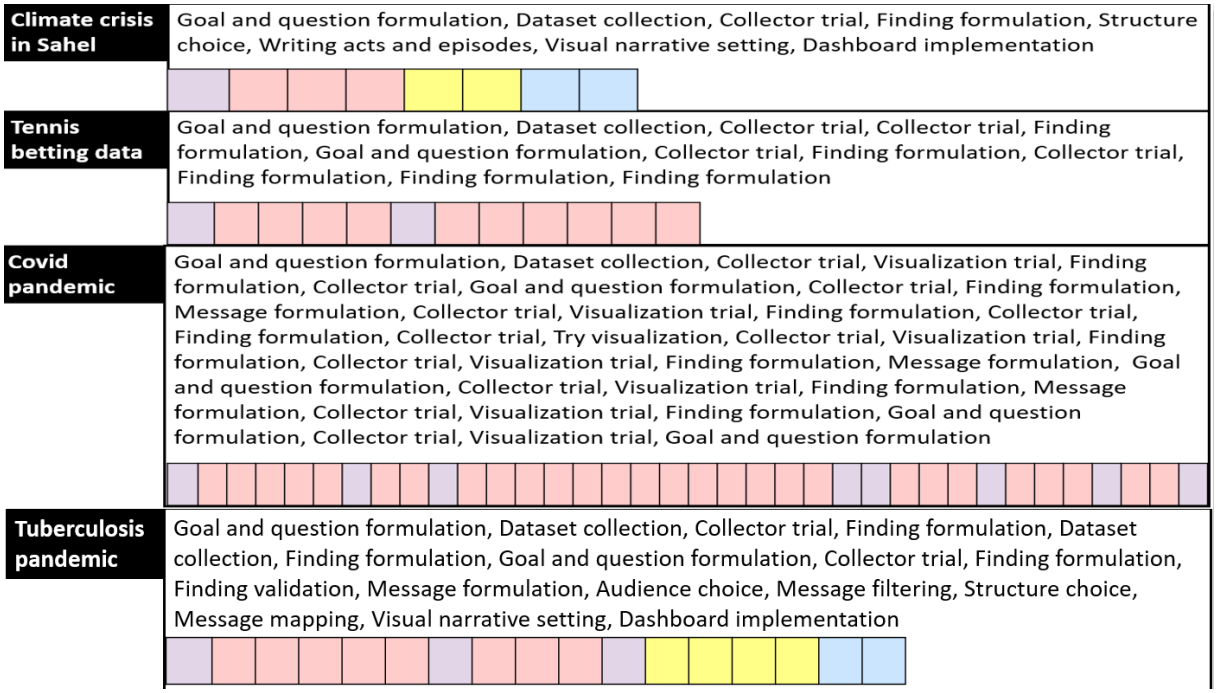


Figure 4.11: The activities of documented processes created by various data narrators with specialized skills.

Figure 4.11 lists the activities performed in the analyzed processes, which are also sketched as a sequence of boxes, colored as the phases of our process.

Results. As a first remark, all the reported activities could be matched to those of our process and the flow between activities is also consistent with our process. In addition, all processes describe many iterations among the initial phases, even if some of them just illustrate some examples of questions and collectors. All of them follow the exploratory scenario. Furthermore, we remark that intentional activities (those of the answer questions phase) are present in all the reported processes.

4.5.4 Importance of the data narrative process

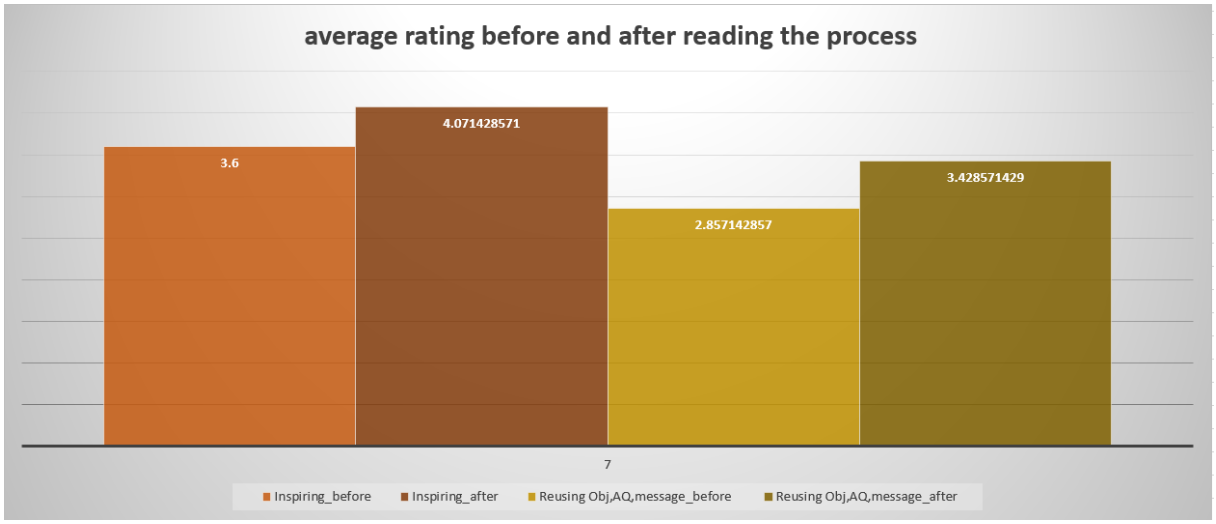


Figure 4.12: Average rating before and after reading the process, based on the criteria of inspiration, and reuse of intentional components.

In this subsection, we conducted a preliminary study via a satisfaction survey and present the methodology used, along with the results obtained in this study.

Methodology. A preliminary study via a satisfaction survey in French (from one to five levels) investigates 15 master's students in Computer Science, who specialize in data analysis. Students received a one-hour tutorial on data narratives, which included the presentation of definitions and examples, the emergence of the conceptual model for data narratives (see Figure 3.3), and the introduction of the typical crafting activities. Students have an experience with data analysis and visualization tools, but none have an experience with data narratives. The survey is broken up into two parts:

- At the first time, students were asked to read a data narrative entitled by *A Race to Adapt: The Climate Crisis in the Sahel* [13]. Then they were given a list of questions to answer.

- At the second time, students were asked to read the process of crafting the data narrative [14] developed by the data journalist who had written the data narrative. After reading the process, students were asked the same set of questions once more.

The survey contains a set of questions based on the given document regarding the level of: being *inspired* to produce another data narrative and *reuse of the objective, analytical questions and messages* adopted in the data narrative. Figure 4.12 illustrates the survey's outcomes.

Results. The survey reveals that students found the reusing of the objective, analytical questions, messages and the structure of the data narrative easier after reading the process. Further, after reading the process, students are more inspired by the data

[13] https://data.humdata.org/visualization/climate-crisis-sahel/?_gl=1*1pnhp3t*_ga*MzE0MjI3MDkzLjE2NzA1OTU0NjY.*_ga_E60ZNX2F68*MTY3MDU5NTQ2NS4xLjAuMTY3MDU5NTQ2NS42MC4wLjA.

[14] https://centre.humdata.org/developing-a-data-story-on-the-climate-crisis-in-the-sahel/

narrative to produce another. *Survey highlights the importance of having a process behind each data narrative. Such a process facilitates the reusing of data narrative components in order to share and reproduce another data narrative.*

4.6 Conclusion

In this Chapter, we proposed a process for crafting data narratives, that covers the whole cycle of data narration, from data exploration to the visual presentation of the 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, it accommodates a wide range of practices observed on the field, via clearly delineated activities, while being well founded upon the conceptual model of the domain detailed in Chapter 3.

The completed studies demonstrated the significance of having a process for crafting data narratives as well as how the proposed process incorporates activities found in other studies. Furthermore, the studies that have been done show that intentional activities have an impact on narrative quality, supporting our claim that intentional aspects are crucial for crafting data narratives.

Chapter 5

Data narrating cube explorations

Albert Einstein said: '*The whole of science is nothing more than a refinement of everyday thinking*'.

Beginning with a broad overview of modeling data narratives and drawing inspiration from the quote cited above, this chapter focuses on refining the data narrative conceptual model accounting for a particular context, data cube exploration. Specifically, we focus on refining the factual and intentional layers of the conceptual model.

Messages and findings, the key concepts of the factual and intentional layers, are carefully studied, taking into account the data narrator's intentions and considering various criteria, such as the data narrator's beliefs and the historical context of data narrative creation.

Contents

5.1 Introduction

The conceptual model introduced in Chapter 3 is designed to identify the key concepts of data narratives across diverse domains. However, it operates within a broad scope, offering a panoramic view rather than delving into the characteristics of each concept such as datasets, collectors, etc. This chapter is dedicated to characterize the conceptual model within the context of data cube explorations. Specifically, we refine the concepts of the factual and intentional layers of the model to provide a more detailed understanding of

these concepts, which will serve as a solid foundation for the subsequent advancement of tools aimed at automating the generation of findings and message within the corresponding factual and intentional layers of data narratives.

This work was initiated by Professor Panos Vassiliadis and PhD candidate Dimos Gkitsakis. This work has continued in collaboration to introduce a specific type of finding called "highlights" [162]. The collaborative efforts are detailed in Sections 5.2 and 5.3. Within this chapter, our exclusive contribution revolves around the presentation of the scoring model for highlights and phenomena, as well as the modeling of the message within the context of data cube exploration. These aspects will be further elaborated upon in the subsequent section, which is referenced as 5.4.

Cubes have considerably contributed to the management and resolution of these complexities by providing a structured method for multidimensional data analysis. This design has the following advantages:

- Schema is clean and follows an archetype *simple* structure;
- Queries are typical analytical queries that allow the analyst to handle vast amounts of data via filtering and aggregation;
- There are no complex operations such as multiple table joins required to work with the data. In addition, the generalization to the realm of in-situ data management, in which data analysts work with basic structured files obtained from arbitrary sources, is relatively simple.

Cubes, in the context of data analysis, refer to a multidimensional representation of data that allows for efficient querying and analysis. Within this context, "dimensions" signify the specific attributes or categories by which data is categorized, "measures" represent the quantitative metrics under examination, and "facts" encapsulate the actual data values themselves. Please note that when referring to elements of data cube exploration, we will use the term "indicator" rather than "measure". The goal of this modification is to ensure easy reading and eliminate any confusion that might arise. We aim to maintain consistency throughout this PhD thesis, and since the concept of "measure" was first introduced in Chapter 3, we have made this change.

Data cubes exploration [61, 126, 161], also known as OLAP (Online Analytical Processing), plays a pivotal role in the world of data-driven decision-making and business intelligence (BI). These technologies are instrumental in enabling organizations to gain deeper insights, make informed decisions, and extract valuable knowledge from vast and complex datasets. Data cube exploration enables users to explore data from multiple perspectives, drill down into details, perform calculations across different dimensions, and others sophisticated operations, including slice-and-dice, roll-ups, pivoting, and more.

Example 1. As an illustration of data cube context, assume we have a database with product sales. The fundamental source of information is a fact table *SalesFact(ProductId, TimeId, CityId, PromotionId, CustomerId, Sales, Costs, SalesUnit, AvgSalesPerUnit, Profit)*, with all the monetary measures expressed in thousands of Euros. Around the aforementioned fact table, we also have several lookup dimension tables, namely *Products, Time, Cities, Promotions, Customers*. This is as typical a Business Intelligence setup as it gets.

The data cube environment proves to be exceptionally well-suited for the refinement of the general conceptual model for data narratives. Unlike the general model, which naturally lacks specific details about the underlying data and the data collection process, data cube environment offers a specialized framework designed to handle data in a unique way. It includes dedicated OLAP operators tailored to analyze multidimensional data, allowing for in-depth exploration and understanding of the dataset's intricacies.

The exponential increase in the number of possible combinations and relationships between dimensions can make the analysis of data in multidimensional spaces inherently complex. Various approaches [33, 38, 59, 61] have emerged to tackle these challenges, including dimensionality reduction, specialized databases, data cleansing, visualization and advanced analytics methods.

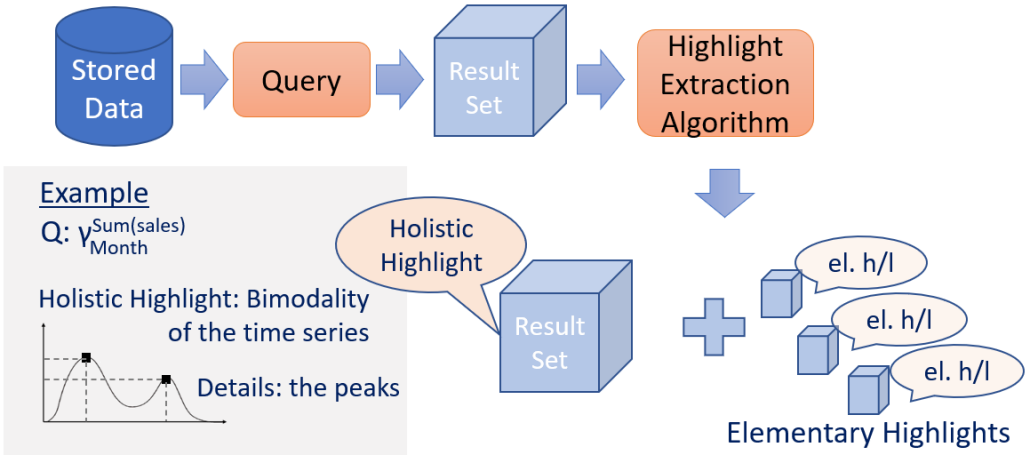


Figure 5.1: Data cubes exploration [162]

In this Chapter, we rely on the approach for data cube exploration illustrated in Figure 5.1. One crucial aspect of this exploration is the extraction of highlights, which provide structured testimony for the existence of a property within a specific dataset. These highlights are automatically tested using a dedicated algorithm and characterized accordingly. They represent a particular type of findings in the data cube context.

Data cube exploration involves the practical application of a set of Highlight Extraction Algorithms to analyze a dataset, typically obtained as the result of a query. These algorithms function much like pattern-matching testers, their primary objective being to determine whether the data conforms to specific patterns or not. To give a couple of concrete examples, possible questions that such algorithms might ask are: (a) is there a bimodality in a time-series produced as a query result, and if yes, which are the peaks?, (b) if we breakdown the total sum of sales by product type, is there a "mega-contributor" product type (and if yes, who is it), with more than 40% of total sales?, (c) assuming we group by total sales per month and product is there any month that systematically outperforms all other months for all products, and if yes, who is it?

An an illustration of Highlight, assume a query that selects the total sum of sales of the product *Wine* for the 2nd quarter of 2023, grouped by month and city. Observe the following *highlights* of the result set:

- The city *Athens* dominates all other cities: for every month, the sales of Athens are higher than the sales of every other city.

- The month *May 2023* dominates all other months: for every city, the sales of May are higher than the sales of other months.
- The city of *Athens* is a mega-contributor to the total sales: the sales of *Athens* are 75% of the total sales.
- If one observes the time-series of the marginal sales per month, there is no trend or seasonality; however there is a unimodality in the time-series: sales rise, reach a peak (in *May*), and then drop.

	Athens	*Rhodes*	*Chania*	*Thera*	
April 2023	500	50	85	80	*715*
May 2023	1000	70	90	120	*1280*
June 2023	600	65	70	70	*805*
	2100	*185*	*245*	*270*	***2800***

Table 5.1: Reference Example

How do we report the results to the user? What we want is a system that derives a *data narrative* that includes (a) the data per se (as E.Tufte has famously emphasized *always show the data*), (b) appropriate visualizations that are automatically produced, and, (c) text narrating the essential elements of the data. The following textual summary reports all the discovered highlights, and groups them around the main *Dimensions* of Athens and May:

> *In terms of geography, Athens dominates all other cities, in every month. In fact, Athens is a mega-contributor to total sales, by contributing 75% of all sales. In terms of time, the progression in time shows a peak in May; in fact, May dominates all other months in terms of total sales. No trend or seasonality were detected.*

This chapter is organized as follows: In Section 5.2, we introduce novel concepts that are necessary for the transition from a general data narrative to the particular field of data cube exploration. In Section 5.3, we introduce the model for highlights, the key concept of the factual layer, discriminating highlights in *Holistic Highlights*, which are properties of the entire set of data being examined, and *Elementary Highlights* which concern individual *Dimensions*, or sets of them, that play a crucial role to *Holistic Highlights*. In Section 5.4, we detail the model for message, the key concept of the intentional layer in the context of data cube exploration. Also, an intrestigness scoring model is proposed taking into account both the intention of the data narrator and the historical context of data narrative crafting, including the conducted exploration, retrieved findings, and other relevant concepts. Throughout the chapter, we will provide an example to help understand the models of highlight and message. Section 5.5 summarizes the chapter's conclusion.

5.2 Refining data narratives within data cube exploration

The key concepts of the factual and intentional layers are "finding" and "message", respectively. In essence, a message acts as the response to an analytical question, backed by one or more findings. These findings are represented as "Highlights" in the context of data cube exploration. An analytical question is transformed into a technical question to be later transformed in the form of an SQL or MDX query for example, a common practice in the context of data cube exploration.

Figure 5.2 provides an overview of the refinement of both the factual and intentional layers within the realm of data cube exploration. This refinement involves an examination of the existing concepts from the general model. Some of these concepts will be retained or refined, while others will give rise to entirely new concepts as they are tailored to the context of data cubes.

All along this Chapter, the color of the models aligns with the color code utilized in the conceptual model for data narratives. Specifically, the use of pink denotes concepts that are associated with the factual layer, encompassing factual aspects. Conversely, the adoption of the purple color signifies concepts that belong to the intentional layer, which encompasses elements related to intentions, goals, and subjective aspects. This harmonization in coloration aids in visually distinguishing concepts based on their inherent layers and characteristics within the model.

This section delves into novel concepts that have not yet been incorporated into the general conceptual model of data narratives. By specifically adapting the concepts of the general model to the field of data cube exploration and taking into account the particular needs of this context, new concepts may arise as we refine the conceptual model in the context of data cubes. Specifically, we introduce the following concepts:

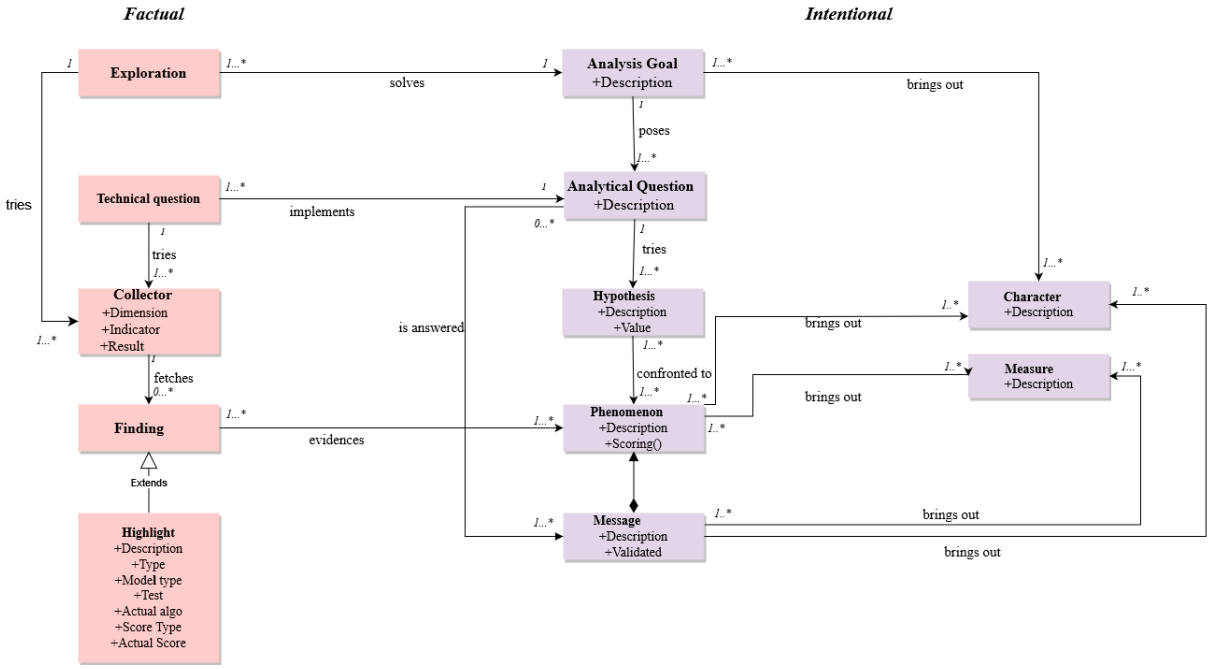


Figure 5.2: A class diagram modeling the factual and intentional layers (Chapter 3) within the context of data cubes explorations

- **User model**. A user model represents the beliefs and the history of actions of the data narrator. It is constructed based on the knowledge and understanding of the data narrator, as well as the past activities, including the formulation of analytical questions, findings, messages and phenomenons.

- **Hypothesis**. A hypothesis suggests a tentative answer or prediction that can be tested through further analysis. As an illustration, consider the following hypothesis: *Students who have regular access to online math tutorials will demonstrate higher academic performance in mathematics compared to those who do not have such access.* In order to answer a hypothesis, the hypothesis is confronted to a rigorous examination and comparison against a specific phenomenon. This process involves analyzing and evaluating the relationship between the hypothesis and the observed phenomenon to determine if there is empirical evidence to support or refute the hypothesis.

- **Phenomenon**. A phenomenon is the algorithmically verified existence of a specific property of internal relationships between the contents of a dataset. For instance, when an algorithm like "Moving Average" detects a property of clients who are increasing their purchases over time, a "Purchase Trend" phenomenon appears in the dataset. Phenomena act as bridges between facts and messages, as they evidence the existence of interesting properties of the data at a higher level of abstraction, in a concise (practically Boolean) way.

- **Highlight**. In the specific context of data cube exploration, a highlight is a particular type of finding, the key concept of the factual layer introduced in Chapter 3. In more details, a highlight is *the structured annotation of the parts of the data that*

make the existence of an interesting phenomenon true with the necessary information that explains why these data are of importance.

- **Technical question**. A technical question is the translation of an analytical question into a technical language, enabling the implementation of data collectors for data narrator with varying technical proficiencies, including SQL, Python, and other related skills. In the context of data cube exploration, we restrict a collector to the set of OLAP operators applied within data cubes. These operators include a range of actions such as aggregation, slicing, dicing, drilling down, and rolling up. A technical question encompasses the necessary instructions and language to formulate a query that addresses the technical question at hand. For instance, consider the following technical question as an illustrative example within the realm of relational algebra. We use the notation provided in [63] to craft an analytical query regarding wine sales in cities across Greece:

  $\gamma_{\text{city, SUM(sale\_amount)}}(\sigma_{\text{product\_name}='Wine'}(sales))$

We also refine the concept of a *message*, initially introduced in Chapter 3. A message is refined as a combination of one or more phenomena that clearly manifests a set of characters and measures. The validity of the message is ensured through various means, such as the involvement of experts or credible sources, to establish its reliability and accuracy.

The next sections details the modeling of highlights and messages within the context of data cube exploration.

5.3 Highlights for data cube exploration

In this section, we will explore the specifics of Holistic and Elementary highlights. Then, in order to illustrate the concept of highlight, we offer a motivating example. Specifics of the highlight model will be discussed later.

In the following Subsections, we will explicitly differentiate between the holistic highlights and the elementary highlights.

5.3.1 Models for Highlights

Based on whether a highlight concerns the entire result set of a collector (like in the case of correlations), or, specific records or values in it (like in the case of mega-contributors), we dichotomize highlights as follows:

- *Holistic Highlight.* A *holistic highlight* is an automatically extracted observation of importance that characterizes a dataset. The existence of a property is tested automatically via an appropriate algorithm, and the result is a model of the data, along with a set of detailed highlights that pertain to individual records, or values of the dataset. A possible textual description of a Holistic Highlight is as follows:

  > The $< HighlightType >$ for $< Indicator >$, tested via $< Algorithm >$, $\{SupportiveRoleText; SupportiveRole\}\star$, fits under the $< Model >$ with $< ScoreType >$ and value $< Score >$.

- *Elementary Highlight.* An *elementary highlight* is a fact produced by the combination of specific Dimensions and Indicators that play an important role in a holistic highlight. A possible textual description of an Elementary Highlight is as follows:

  > The $< DimensionalAttribute > = < DimensionalValue >$ with the value $< Indicator > = < IndicatorValue >$ serves as $< HighlightType >$ with a $< ScoreType > = < Score >$.

The metamodel for highlights within multidimensional space is depicted in Figure 5.3. In the following, we will begin by detailing the holistic highlight and enumerating its distinct characteristics. Subsequently, we will explore the elementary highlight in-depth, along with its specific characteristics.

Additional concrete examples illustrating the concept of Holistic Highlights and Elementary Highlights are instantiated and displayed in the form of Tables in Appendix 1.5. These examples further exemplify the distinction between the two types of highlights and provide visual representations of how they manifest in real-world data scenarios.

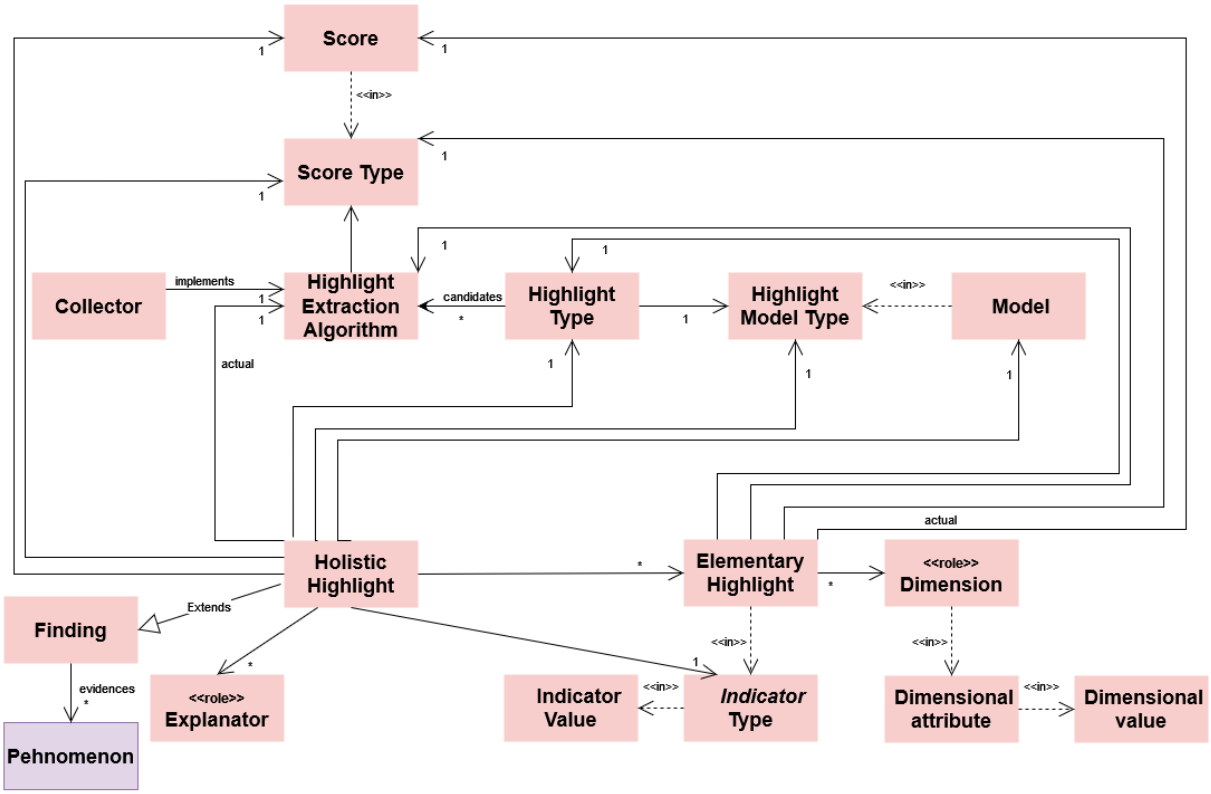


Figure 5.3: The metamodel for highlights in the context of data cube exploration [162]

5.3.1.1 Holistic highlights

A **Holistic Highlight** is a structured testimony for the existence of a property over a specific dataset that is automatically tested via a dedicated algorithm and characterized accordingly. Structurally, a holistic highlight is characterized by the following properties:

- The *Highlight Type* defines the family to which the highlight belongs (e.g., Correlation, Trend, Seasonality, Unimodality, Bimodality, etc.)
- The *Highlight Model Type* defines the type of the result of the investigation of the existence of the property or pattern over the result set. Fundamentally, it defines the domain of values that the assessment can produce. For example, when assessing the correlation of two Indicator Types the Model can be any of *Positively Significant*, *Insignificant*, *Moderately Negatively Significant*, etc.
- The *Model* of the highlight defines the state of the data with respect to the Highlight Type under investigation. Practically, this is achieved via the specific value from the domain of the Model Type that pertains to the result set being characterized.
- The test of the property is performed via one of the many candidate algorithms that exist for the same Highlight Type. For example, Correlation has several candidate algorithms: Pearson, Spearman, Kendall, etc. However, every highlight is produced via the execution of a specific algorithm out of the set of candidate ones, which we refer to as *Actual Algorithm*. The actual algorithm uniquely determines the Model Type of the highlight.
- Moreover, the strength of the property is characterized by a *Score Type* and an actual *Score*. The *Score Type* is the domain for the actual score – e.g., a Score Type *p-value* can have an actual *Score* value of 0.3.
- Typically, the highlight has an *Indicator Type* of the result set that is tested for the existence of a property. For example, the existence of a Bimodality is tested over a certain Indicator Type, say $SumSales$. Indicator Value is the actual value for Indicator Type.
- The indicator alone is not sufficient to give all the information for the property. For example, unimodality of an indicator has to be produced with respect to a certain time attribute (in a typical BI dataset, a basic cube can have several such attributes, $OrderDate$, $DispatchDate$, $ArrivalDate$ – therefore the testing of a modality needs to be done with respect to one of them). As another example, a certain indicator demonstrates strong correlation with respect to another indicator. Thus, a set of Explanators is used to accompany the indicator of the highlight in order to fully specify the existence of a property.
- Last but not least, a holistic highlight can include a set of details in the form of Elementary Highlights.

The role of roles. We make a deviation here to introduce a new stereotype, *Role*. Roles are classifiers, whose operation is to annotate the relationships of highlights to participating entities with extra information. A Role annotates a class with specific attributes, specifically: (a) a *name*, (b) an accompanying *textual description*. This allows a highlight which for example comes with several Dimensions playing a role in it, to be able to discriminate which Dimension plays which role. Thus, within the context of each highlight, the name of a role is unique.

For example; the "Store Locations" dimension could be assigned the role of "Geographical Impact".

An example of a textual description of a Holistic Highlight is as follows:

> The $< Trend >$ for $< MonthlySales >$, tested via $< LinearRegression >$, fits under the $< StrongpositiveTrend >$ with an $< R - squared >$ and value $< 0.85 >$.

5.3.1.2 Elementary Highlight

An Elementary Highlight is a fact produced by the combination of specific Dimensions and Indicator Values that play an important role in a holistic highlight. Structurally, an Elementary Highlight is characterized by the following properties:

- The *Highlight Type* defines the family to which the highlight belongs (e.g., a peak in a unimodality distribution, a part of the top-k facts set for an indicator, a mega-contributor fact over its peer facts for the breakdown of an aggregate value, etc.)
- A *Dimension* is a Role, carrying a distinctive name and text that characterize its functionality with respect to the highlight.
  To illustrate, suppose we have a query that aims to determine the total sales per month and city for a retail business. In this context, we can identify two dimensions: the "Time" dimension and the "Location" dimension.
- The *Dimensional attribute* refers to a specific characteristic or property within a dimension. It helps to categorize, segment, or analyze data along that dimension. In our scenario, the attributes "Month" and "City" are considered Dimensional Attributes.
- The *Dimensional value* represents a specific data point within a Dimensional Attribute. It is a concrete data value associated with an attribute. In our example, "April 2022" and "Athens" are Dimensional Values.
- The *Indicator Type* refers to a specific metric that is used to quantify and analyze the information contained within a data cube. For instance, "SumSales" can be considered a "Indicator Type" because it represents the total sales figure, which is a type of measurement.
- The *Indicator Value* is the specific numerical quantity or value associated with a particular data point within the dataset. It represents the actual measurement for a specific indicator. For example, when you have the sum of sales in Athens for April 2022 as $50,000$, this $50,000$ is the "Indicator Value" because it is the specific numeric result of the measurement for that particular Indicator Type.
- As with *Holistic Highlights*, *Elementary Highlights* are also related to the execution of the actual algorithm that produced them via the respective *Model* for which they perform an illuminating role.

- Similarly, each Elementary Highlight has a *Score Type* and an actual *Score*, to characterize the strength of the highlight. For example, this can be the rank for top-k facts, the percentage of a total sum for a mega-contributor, the percentage of dominated peers of a character for a peer-dominator, etc.

An example of the textual description of an Elementary Highlight is as follows:

> The $< Month > = < April >$ with the value $< TotalSales > = < 50,000 >$ serves as $< PeakinaUnimodalityDistribution >$ with a $< Rank > = < 1 >$.

5.4 Messages for data cube exploration

Message is a crucial concept in data narratives. It serves as an essential connection between the data narrator's intentions, data analysis and narratives. A message emerges from the data, supported by findings, interpreted by the narrator, structured into a coherent narrative, and then communicated to the audience visually.

We have modeled the main components of a message such as highlights and a scoring model for highlight in the context of data cubes exploration, the necessity for modeling "Messages" becomes conspicuously clear within this domain. The modeling of message should take into account the specificities arising from the context of cube exploration. In this section, we will model message and its related concepts within such context.

5.4.1 Model for messages and phenomena

A message acts as the narrative's core, bringing forth key characters and measures. This message is constructed from phenomena, which are concrete data-based observations. This intricate relationship between message, phenomena, and hypotheses ensures that the resulting narrative is robust and well-supported by the underlying data.

Figure 5.2 depicts a class diagram modeling the refinement of the factual and intentional layers of the general model. In the following, our attention will be directed towards providing an in-depth exploration of the "message" concept and their interconnected concepts. In order to provide a more detailed viewpoint, Figure 5.4 focuses in on the model's refinement with respect to the "message" concept. Concurrently, Figure 5.5 focuses on the concept "user model" and offers a more in-depth analysis of this concept of the refinement model.

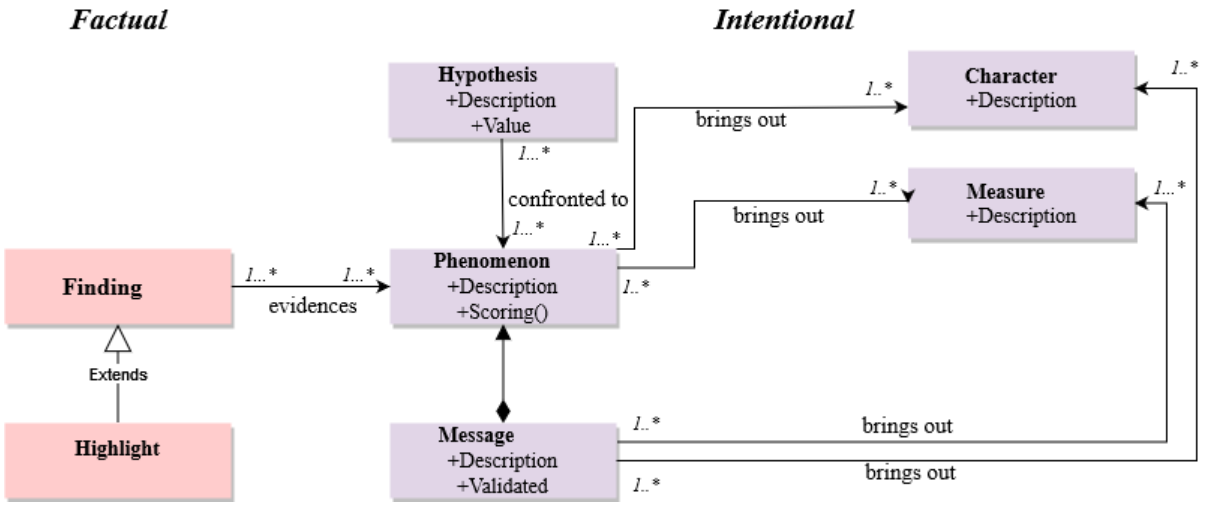


Figure 5.4: A class diagram focusing on *message* within the context of data cubes explorations

Note that certain relations and cardinalities may vary from those shown in Figure 5.6 since we have assumed that the introduced scoring model is general and separate from this.

To provide an illustration and distinguish among phenomena, highlights and messages, we will present an example for each of these concepts. A phenomenon can be "Athens has higher sales than other cities, verified via an ANOVA test and supported by a highlight like " The Dominance for Sales, tested through Comparative Analysis of monthly sales data, confirms that Athens dominates all other cities, with a p-value of 0.3" ". A message is composed of one or more phenomena in this form: "Athens not only maintains higher sales than other cities but also exhibits varying degrees of dominance in different months, implying a distinct seasonal sales pattern".

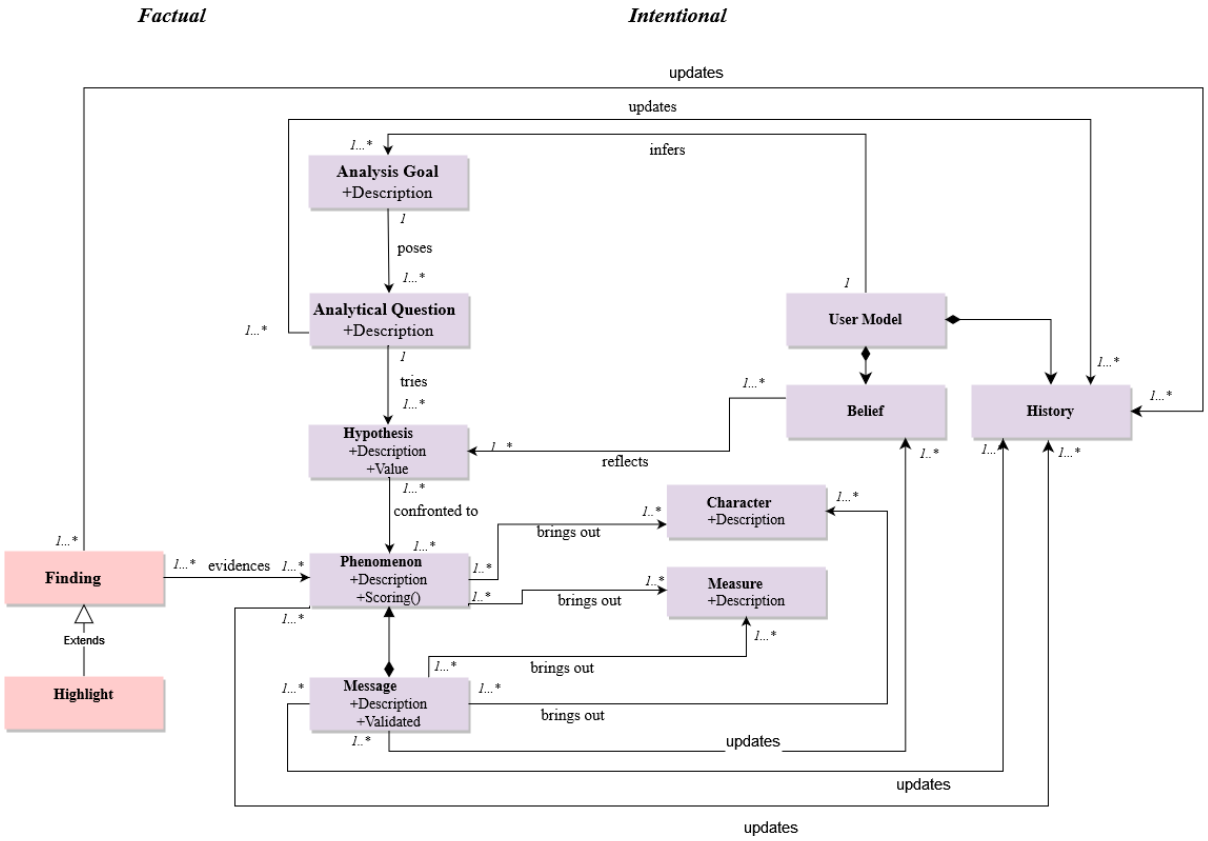


Figure 5.5: A class diagram focusing on *user model* within the context of data cubes explorations

Transforming hypotheses into verified messages The analysis goal of the data narrative is solved by an exploration that aims to retrieve pertinent and informative messages supported by data. This goal reveals one or more analytical questions, each representing a specific aspect of the goal. These analytical questions, in turn, lead to the development of one or more hypotheses, serving as provisional answers containing a property that is relevant to the specific question. This hypothesis must be verified through the data exploration.

To test these hypotheses, one or several technical questions can be formulated. These technical questions are instrumental in transforming the intentional aspect into an approximate factual aspect, bridging the gap between the intention of the data narrator and the factual exploration of data. The concept of a technical question plays a vital role in facilitating the automation and transition between the data narrator's intention and the factual data analysis, acting as a translator bridge between the two.

To implement a technical question, one or several collectors are tried to explore data. Highlights are not merely the result obtained from the collector; they also include evidence and importance score that aid in understanding and comparing the results. The model for scoring highlight is detailed in Subsection 5.4.2.

One or several findings can either confirm or refute a hypothesis. A validation of a hypothesis is a prerequisite for it to be considered a phenomenon.

The model for scoring phenomenon is detailed in Subsection 5.4.2. A message is composed of one or several phenomena, which are validated by an expert to ensure that only

validated messages are disseminated, thus avoiding the spread of unverified information.

5.4.2 Intrestigness scoring model

Once highlights are generated and phenomena are derived, the selection of the most intresting ones arises. The goal of an intrestigness scoring model is to assign scores and select a subset of highlights and phenomena among several available highlights and phenomena.

To address this, we employ an intrestigness model based on data narrator intention and a well-defined scoring framework [69, 107] in data cube environment.

The intrestigness scoring model differs from the concept score introduced in Figure 5.3, which is used to evaluate individual highlights.

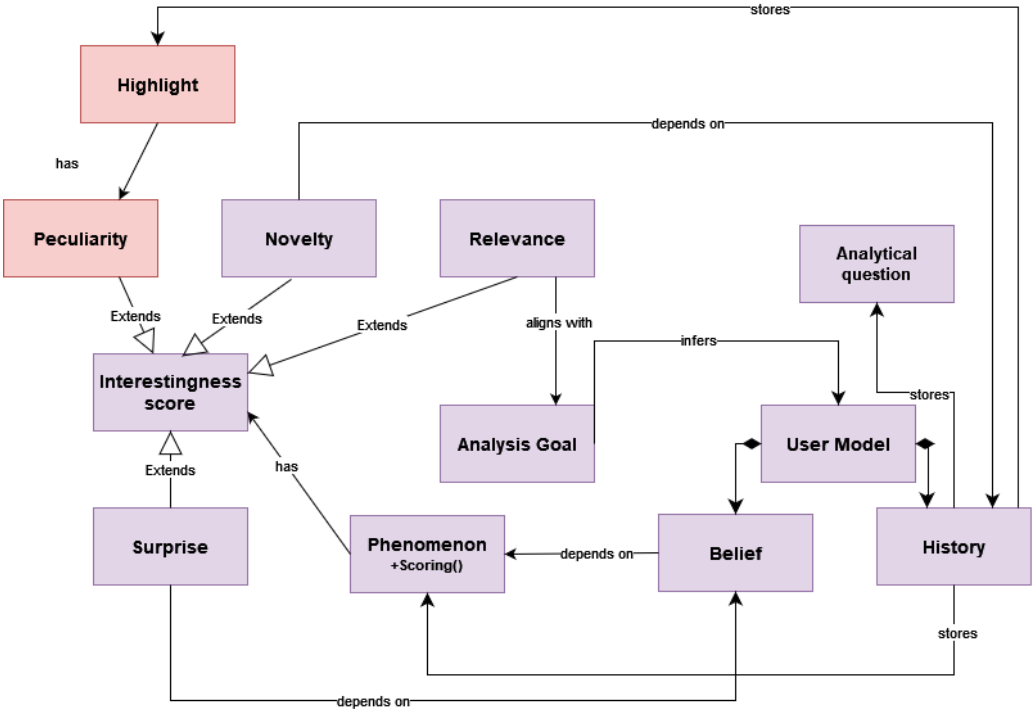


Figure 5.6: Model for scoring findings and phenomenons

Figure 5.6 depicts the model which has been specifically crafted to assess and assign scores to highlights and phenomena within the context of data cube exploration.

This model draws its foundation from the concept of "interestingness score" as elaborated in [107]. It is tailored to the computation of an interestingness of a particular cell derived from a multidimensional data cube. This approach incorporates considerations of human behavior and preferences as integral factors in the process of quantifying the significance of these data elements. The four aspects of interestingness assessment are used to evaluate queries within the context of data cubes [69].

In details, Marcel et al. [107] defined a degree attributed to a piece of information regarding the interest it generates. This degree is characterized by four aspects:

- **Relevance**: the extent to which the results of the query are related to the overall goals of the data narrator.

- **Surprise**: the extent to which the result of the query contradicts, revises, updates the data narrators prior beliefs.
- **Novelty**: the extent to which the result of the query presented to the data narrator is new, and previously unseen to them.
- **Peculiarity**: the extent to which the result of the query is different, and not in accordance with the previous data of the session or history.

Highlights are, by nature, related the factual layer and more concretely to the data. Highlights are not directly related to the user model. Therefore, their score is inherently a score of peculiarity that depends solely on the data.

On the other hand, the scoring of *phenomena* is closely tied to the user model. It is based on an "interestingness score", which takes into account three factors contributing to its level of interest. A scoring function is assigned to score phenomena based on the following factors: (i) the relevance of the phenomenon to the goal of the data narrative, (ii) the novelty of the phenomenon compared to the information history of the data narrator, and (iii) the level of surprise it evokes in relation to the data narrator's beliefs. The data narrator's beliefs are derived from the information present in the data narrator's mind. The analysis goal may be inferred from the user model. For instance, by analyzing a user's past behavior and preferences, a system can make assumptions about the user's current goals. The user's history comprises a collection of phenomena, messages, findings, and goals used in the creation of the data narrative.

5.4.3 Example of the refined model

In this example, we will extend the scenario introduced in Example 1, which was described in Section 4.1 and pertained to product sales. This extension will offer more concrete instances of the concepts within the models. The data narrator aims to explore this database to craft a data narrative.

Analysis goal	Poses an **analytical question**	tries a **hypothesis**	confronted to **phenomenon**	brings out **characters**	brings out **measures**
-Description: analyze the wine sales market in Greece	**AQ1:** -**Description**: How do wine sales evaluate in different cities in Greece?	**HYP1:** -**Description**: Some cities have higher sales of wine in Greece than other cities -**Rejected**: NO	**PH1:** -**Description**: Athens have the higher sales than other cities, verified via ANOVA test and evidenced by **H1**	-**CH1:** Athens -**CH2**: sales -**CH3:** cities	-**M1**: Sum (Sales)

Figure 5.7: Example of message (1)

The data narrator's goal is to examine the Greece's wine sales market. This goal is formulated based on the data narrator's belief that the wine sales vary across different cities within the country. This goal brings out characters as Greece, wine and sales. To

delve deeper into the subject, an analytical question arises: "How do wine sales evaluate in different cities in Greece?". This question highlights the same characters mentioned in the goal. This question also tries a hypothesis suggesting that certain cities in Greece have higher wine sales than others. This hypothesis is subsequently accepted, as it aligns with a phenomenon indicating that Athens has the highest wine sales in Greece verified via ANOVA test and evidenced by one highlight. The phenomenon brings out the characters and includes one measure regarding the total sales. This phenomenon is scored high based on the relevance of the phenomenon to the goal of the data narrative, the novelty of the phenomenon compared to the information history of the data narrator, and the level of surprise it evokes in relation to the data narrator's beliefs. Furthermore, the phenomenon will become an integral part of a newly crafted message.

Technical question	tries **collector**	fetches for **highlight**	evidences **phenomenon**	composes **message**
TQ1: **-π (city, SUM (sale_amount) (σ (product_name ='Wine') GROUP BY 'city'** **-Implemented in SQL**	**C1:** -**Query**: SELECT city, SUM (sale_amount) as total sales FROM sales WHERE product_name ='Wine' GROUP BY 'city' -**Dimension**: city -**Indicator**: SUM(sale_amount) -**Result**: |Athens|$10,000| |Rhodes|$5,500| |Chania |$3,200|	**H1**: -**Description**: The Dominance for Sales, tested via Comparative Analysis of monthly sales data, affirms that the city Athens dominates all other cities, with a p-value of 0.3 -**Type**: Dominance - **Model Type**: Statistical test **-Model**: Dominance of Athens over other cities -**Algorithm**: ANOVA -**Score Type:** p-value -**Actual Score**: 0.3 -**Elementary highlight**: The character set {Athens} with the value Sales serves as Dominance Highlight with a Score Type = Percentage Dominance and a Score Value = 20%	**PH1:** -**Description**: Athens has the higher sales than other cities, verified via ANOVA test and evidenced by **H1** **PH2:** -**Description**: Athens dominates all other cities in different months with varying degrees, indicating a seasonal variation in sales, verified via ANOVA test and evidenced by **H1** **+Scoring ():** **{Novelty, relevance, surprise}**	**M1**: -**Description**: Athens not only maintains higher sales than other cities, but also exhibits varying degrees of dominance in different months, implying a distinct seasonal sales pattern. -**Validated**: By expert

Figure 5.8: Example of message (2)

We explain now, how the highlight H1 was obtained. An exploration is launched to solve the goal and a technical question is prepared in relational algebra notation to answer the analytical question by exploring data cube. A technical question is then transformed and implemented as a collector in the form of an SQL query. A collector has several parameters like the dimension, indicator and the result of the query. Collector is implemented to fetch for highlight. A highlight is derived from the analysis showing that the dominance for sales, tested via comparative analysis of monthly sales data, affirms that the city Athens dominates all other cities, with a p-value of 0.3. This p-value score indicates a significant distinction from other highlights that is not evident in this example.

User model	is composed by **belief**	and **history**	Scores highlight	Scores phenomenon
-**Goal**: analyze the wine sales market in Greece	-**Description**: sales of wine are not uniform in Greece's country	-**M1** -**Ph1** -**Ph2** -**H1** -**AQ1**	-Peculiarity: High (indicating a significant distinction)	-Novelty: Moderate (as this specific dominance is moderate) -Relevance: High (as it directly aligns with the user's goal) -Surprise: Moderate (the dominance is somewhat surprising but not entirely unexpected)

Figure 5.9: Example of message (3)

The user model brings around the goal and consists of three key elements: the analytical questions asked during the creation of the data narrative, the data narrator's beliefs regarding the uneven distribution of wine sales in Greece, and the data narrator history, including analytical questions, phenomena, messages, and findings derived from the data narrative. The user model is updated once an analytical question, phenomena, message, or finding is obtained.

5.5 Conclusion

In this chapter, we have focused on refining the data narrative model in the context of data cube exploration. We focused on modeling the factual and intentional layers of the conceptual model for data narratives. Messages and findings the key concepts of the factual and intentional layers, are carefully crafted, taking into account the data narrator's intentions and considering various criteria, such as the data narrator's beliefs and the historical context of data narrative creation. We've introduced models that serve two critical purposes: (a) providing an organized structure for modeling and (b) enhancing the clarity of data narrative concepts in the context of data cube exploration. These models refine the concepts of factual and intentional layer within the context of data cube exploration. Furthermore, we proposed a scoring model reflecting the intention of the data narrator within the context of data cube exploration.

Chapter 6

Conclusion

Narratives have distinguished themselves as powerful tools of communication throughout history, able to shape perceptions and understanding. However, their power grows even deeper when these narratives are supported by data. By incorporating narratives techniques, data visualizations, data exploration techniques and other visual elements, data narration offers a powerful way to convey data-driven information in an engaging and informative manner.

This thesis explored the complex domain of data narratives, leading to the development of a framework for crafting data narratives. In a world characterized by a tsunami of data, this framework arises as a guiding principle, shedding light on the concepts and process of crafting data narratives. We navigated the landscape of crafting data narratives and we were confronted with a plethora of research questions that illuminate the path ahead. These questions were answered directly in this thesis: (i) How can data narratives be effectively defined and modeled, considering the novelty and heterogeneity of the data narratives domain? (ii) What are the key components of a framework for constructing data narratives? (iii) What are the primary activities involved in the process of crafting data narratives? (iv) How to account data narration for the particular and typical case of data cube exploration?

6.1 Contributions

In response to the fundamental research questions that guided our research, we accomplished three significant milestones:

- **A conceptual model for data narratives**: we proposed a conceptual model for data narratives that serves as a foundational model for understanding the core concepts within the field. This model facilitates understanding, sharing, and reuse of data narratives, considering the lack of existing data narrative modeling. This model provides a structured, principled definition of the key concepts of the domain, along with their relationships, and clarifies their role and usage.

- **A process for crafting data narratives**: we proposed a process encompassing multiple phases with a flow diagram illustrating the diverse activities involved in data narration, including data analysis, extraction of relevant messages, structuring

of messages into a coherent narrative, and visual representation. This process provides a methodological guidance and facilitates process support through the use of tools, regarding the lack of integrated tools covering the whole crafting process and recommending actions to less-experienced narrators.

- **Illustrating a typical use case of data narrative in the context of data cube exploration**: we refined the data narrative conceptual model accounting for data cube exploration as a specific case of data narration. We modeled the factual and intentional layers of the proposed conceptual model for data narratives, with a particular emphasis on modeling the answer to analytical questions through messages supported by highlights, a specific type of findings retrieved from the data in the context of data cube exploration. By considering various elements such as data cube environment, belief of data narrator and the history of data narrative crafting, the research refined modeling the factual and intentional layer of the proposed conceptual model within the context of data cube exploration.

The developed conceptual model, the defined crafting process, and the emphasis on message modeling contribute to advancing the field and providing valuable insights for researchers and practitioners in the realm of data narratives. This research journey has led us to conceptualize and present a framework that is specifically tailored to the complex art of data narrative crafting. This framework not only incorporates the fundamental concepts of the domain but also provides a comprehensive guide that includes the dynamic workflow and complex crafting activities. The present research doesn't just rely on the body of research that has already been done on data exploration, narratives, and visualization; rather, it makes significant use of real-world data journalism and data science experiences to create data narratives. This combines theoretical understanding with real-world expertise from people who are actively involved in creating data narratives. This research departed from the generic modeling of data narratives, delving deeply into modeling data narratives within the context of data cube exploration.

Several important insights are provided by the carried out studies. Firstly, they illustrate the practicality and applicability of the conceptual model through manual analysis and a proof-of-concept implementation in the context of data narratives. Secondly, they underscore the importance of having a structured process for crafting data narratives and how the proposed process incorporates activities found in other research. Additionally, these studies demonstrate that intentional activities significantly influence narrative quality, thus reinforcing the argument that intentional elements play a crucial role in the art of crafting data narratives.

6.2 Perspectives

Moving towards the frontier of potential perspectives, we believe that these models, static and dynamic, can serve as a stepping stone for future research in the area of data narration, both in the short and long term.

6.2.1 Short term

In short term, works to enrich the conceptual model by adding semantic aspects, automating tasks and activities for crafting data narratives, especially with regard to intentional tasks, will be a focal point of our research.

Semantic aspects. An investigation into how to enrich the conceptual model of data narrative can be useful to facilitate understanding, organizing, and analyzing data within a general or specific domain. For example, consider the inclusion of multi domain ontologies [141] and hierarchies [62]. These semantic aspects can play a crucial role in ensuring that the narrative accurately reflects the underlying subject matter. Through the use of ontologies and hierarchies, data narratives can achieve improved semantic clarity, facilitating not only the exploration of data but also the coherent structuring of the narrative plot. Essentially, these semantic aspects add depth and precision to both the factual and structural layers, facilitating the data narrative crafting.

Automating data narration. This framework for crafting data narratives can be the basis for implementing tools for guiding the data narrator throughout the process as well as automating tedious or complex tasks. The development of automated tools should employ natural language processing techniques [99] facilitating the automatic generation of data narrative concepts. Template-based generation technique [99] is a natural language generation technique in which predefined templates or syntaxes are filled with the appropriate content. It is often used when specific, non-random text needs to be created. The template-based generation technique offers an automated approach to create messages or episodes. This method involves the utilization of predefined templates filled with relevant content to produce content automatically. It is particularly advantageous when there is a requirement to generate specific and structured messages or episodes.

The development of automated tools bridge the gap between advanced analytical tools and non-expert users, facilitating more effective communication across various domains [58].

Benchmarks. Another challenge in the area of developing data narratives lies in benchmarking. This benchmarking is not just about evaluating the final data narrative; It includes all the steps required for its construction. Benchmarking aims to measure the effectiveness and efficiency of each phase of the narrative development process, from data analysis to message creation, structuring and presentation. Establishing meaningful benchmarks for such a multifaceted process is essential to ensuring the quality and usefulness of data narratives, making it a complex but crucial aspect of narrative development in the data narrative landscape.

An additional avenue of research involves the creation of comprehensive evaluation metrics to assess the effectiveness of data narratives. This might include measures for user engagement [95], comprehension [142], emotional impact [87], and retention [43]. Establishing these metrics would enable researchers and practitioners to objectively measure the success of different narrative approaches.

Focus on intentional aspects. We particularly insist on the importance of the intentional phase of the crafting activities. Activities in this phase (e.g., message formulation, message validation) are likely to be the most difficult to automate. Addressing the current gap in incorporating user intention modeling into data narrative crafting necessitates dedicated research efforts to ensure the continuous integration of human in-

volvement. This research was based on the fact that, in order for a data narrative to achieve its objective, an efficient understanding of target audience's interests, expertise and objectives must be attained by collecting user feedback and preferences. Researchers can establish a framework in which data driven insights are seamlessly matched with human understanding, thus enhancing the relevance and impact of information stories by investigating ways to capture user feedback and intentions. By investigating methods for capturing user feedback [85] and intent [180], researchers can establish a framework that seamlessly aligns data-driven insights with human comprehension, thereby enhancing the relevance and impact of the data narratives.

6.2.2 Long term

In long term, there are additional directions to explore, including efforts to enhance the presentational layer, extending the model's applicability beyond multidimensional spaces to cover general data spaces, introducing novel evaluation metrics, and expanding increased collaboration across interdisciplinary teams.

Advanced visualisation. As technology evolves, we are convinced that new opportunities for analyzing, visualizing and presenting data arise through machine learning and 3D representation [78, 104]. Additionally, by delving into emerging technologies like virtual reality (VR) and augmented reality (AR), the current framework can be expanded to enrich data narratives. VR and AR offer immersive, interactive experiences that facilitate user comprehension and engagement with intricate information [88, 153].

VR and AR can provide a three-dimensional, immersive environment for data exploration. In the factual layer, this can greatly enhance the analysis process. Data points can be visualized in three dimensions, allowing data analysts to identify patterns and trends that may be less apparent in traditional 2D data representations. In the presentational layer, VR and AR can create immersive data narratives. Instead of traditional static charts and graphs, data can be presented in three-dimensional spaces, allowing users to explore data as if they were inside the visualizations. This enhances data storytelling by making it more engaging and memorable.

Moreover, the incorporation of an interactive component, as highlighted in the work by Sarikaya et al. [146], in the form of an additional layer introduces a pivotal aspect to the narrative framework. In doing so, VR and AR technologies offer a transformative potential to enhance both the factual and presentational layers of data narratives. They provide new ways to explore data and create engaging, interactive data narratives that are not only informative but also memorable and immersive. Adding interactivity not only allows readers to engage more deeply with the narrative, but also allows data narrators to take a collaborative approach, resulting in more immersive narratives.

Generalizing the Scoring Model: The scoring model that was first presented in the context of data cube exploration is amenable to generalization because its underlying principles do not change regardless of the particular features of the data space. This indicates that the model is robust and versatile, as its applicability transcends the confines of any specific data domain.

Collaboration within interdisciplinary teams. Finally, expanding collaboration within interdisciplinary teams will enhance our understanding of how individuals from different backgrounds can contribute to the creation of comprehensive data narratives.

As an illustration, engage in collaborative efforts with professionals from a wide range of domains, including data science, design, psychology, and communication, to construct a comprehensive framework. Drawing insights from multiple disciplines could lead to a holistic approach that addresses various aspects of crafting effective data narratives. By encouraging interdisciplinary collaboration in future research initiatives, we may uncover new strategies for successful data-driven narrative.

Appendix A

1.1 Data narrative examples

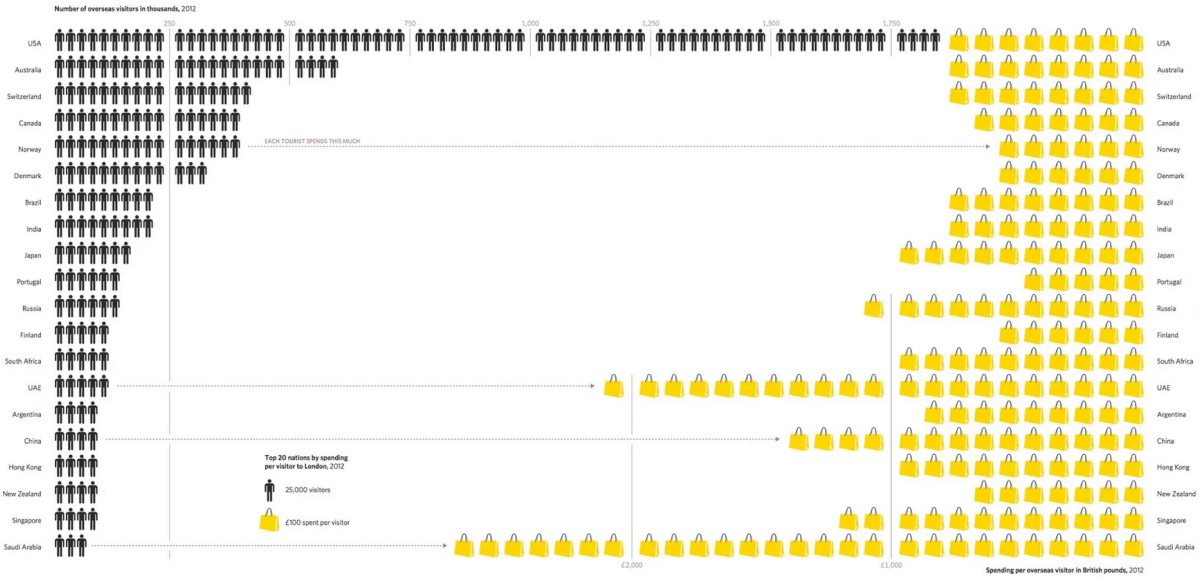


Figure A.1: Top 20 countries in terms of visitor spending in London in 2012 published by the guardian

Figure A.1[1] is an illustration of data narratives, represented in the form of an infographic, pertaining to " The amount spent in London by visitors from various countries". This example conveys multiple messages via brief texts and a single visualization. These messages may include: (i) top 20 countries in terms of visitor spending in London in 2012; (ii) number of international visitors to London in 2012; and (iii) the amount spent by international visitors in London. Various entities were represented by simple symbols in this visualization's elaboration. For example, the sticker on the man represents 25,000 visitors, while each bag signifies that each visitor spent £100.

Figure A.2: Data narrative about Brexit by the numbers: Who voted to leave the EU?

[1]https://www.theguardian.com/cities/gallery/2014/oct/28/london-life-mapped-data-visualisation-graphics#img-6

Figure A.2[2] depicts another example of data narratives. Sky News, a British free-to-air TV news station and organization, crafted a data narrative to explain what would happen if the UK left the European Union. Sky News aimed to shed light on the impact of Brexit during the debate's ambiguity by presenting a clear and straightforward representation of the data.

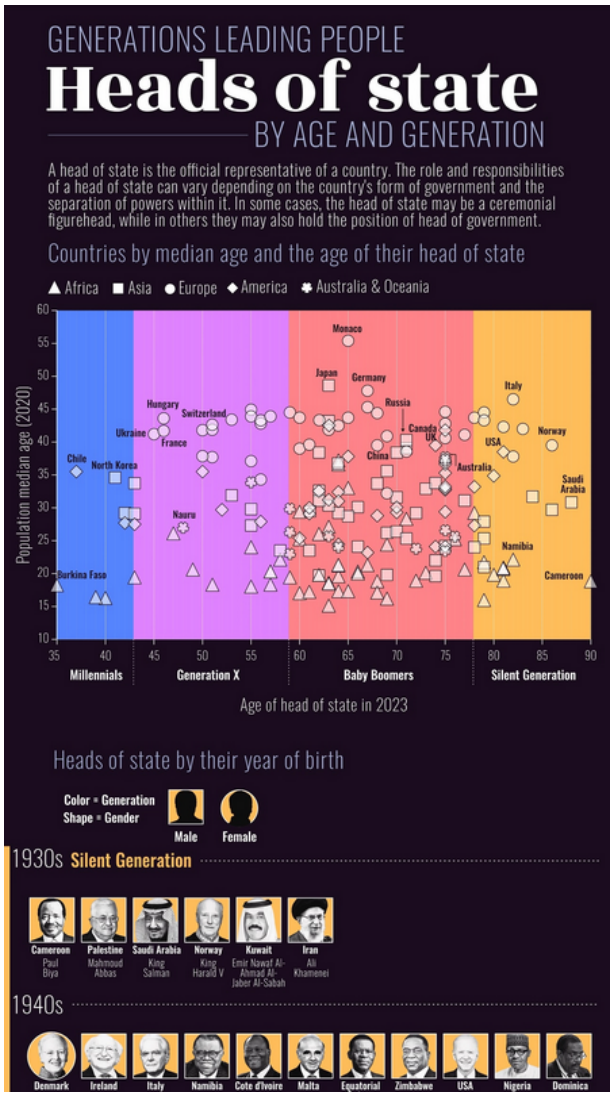


Figure A.3: Heads of state by age and generation

Figure A.3[3] depicts a data narrative in the form of an infographic that comprises both text and visualizations, conveying information regarding the "Ages and generation of heads of state in countries around the world". The conveyed messages may include: (i) the median age and the age of the head of state of countries around the world; (ii) the heads of state's details, including their name, photo, and country, segregated based on the generation to which they belong; (iii) the identification of the oldest and youngest heads of state; (iv) the dominate generation the world's state leadership roles; and (v) the country where the generation "Gen X" takes the lead.

[2]https://news.sky.com/story/better-for-brexit-how-uk-has-changed-since-leave-vote-11920143
[3]https://www.visualcapitalist.com/cp/visualized-heads-of-state-each-country-by-age/

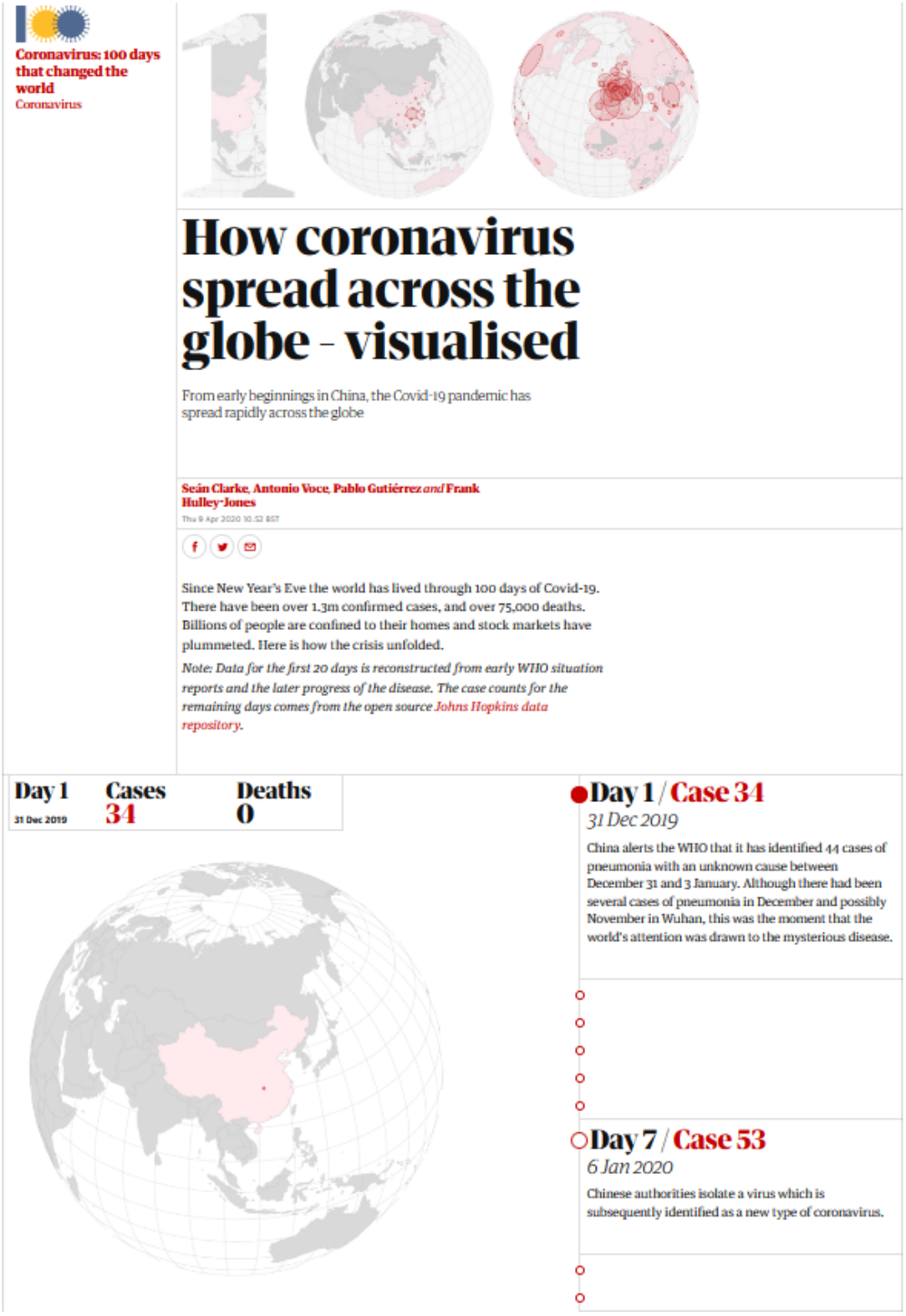


Figure A.4: How coronavirus spread across the globe published by the guardian

Figure A.4[4] illustrates the situation of the coronavirus on a global scale. It relies on an interactive visualization that represents the number of cases and deaths for the

[4]https://www.theguardian.com/world/ng-interactive/2020/apr/09/how-coronavirus-spread-across-the-globe-visualised

reader-selected day.

Figure A.5: A data story about the climate crisis in the Sahel

Figure A.5[5] illustrated a new structure of a data story in the form of an interactive visualization about the climate crisis in the Sahel. In more details, a Chadian girl's daily journey to collect water illustrates how the climate crisis is affecting her community. Rather than moving from one location to another on a map, the data narrator uses images and data visualizations to guide the reader through the journey of the protagonist.

Figure A.6: A data story about the Jews of Lebanon

Although data experts may be inclined to include numerous charts and visualizations,

[5]https://data.humdata.org/visualization/climate-crisis-sahel/

the most effective content teams recognize the importance of utilizing data visualization to serve the overarching narrative. An exemplary instance of this approach is demonstrated in Figure A.6[6], a data story about The Jews of Lebanon published by Arab News, which chronicles the history of a minority community in Lebanon, often marked by tragedy. Despite having access to a wealth of demographic data, the team opted for a simple yet powerful approach, using the size of the Jewish community in raw numbers to emphasize the story's key points. By incorporating data visualization in a strategic manner, the content team was able to enhance the storytelling aspect of their narrative, making it all the more engaging and memorable.

Figure A.7: A data story about New Zealand Labour Party

Figure A.7[7] is one noteworthy example of data narrative published by New Zealand media company Stuff. Data narrator produced a series of interactive bar charts, line charts, and maps to demonstrate the magnitude of the triumph. Although the story utilized some CSS and HTML, the charts themselves were static illustrations, with the interactivity activated as readers scrolled through the content. The visualizations effectively harnessed the power of data to showcase the breadth and scale of the historic event.

[6]https://www.arabnews.com/JewsOfLebanon
[7]https://interactives.stuff.co.nz/2020/10/election-2020-results-analysis-labour-day/

1.2 Towards research papers for implementing data narrative concepts

This section aims to translate the concepts of the data narrative model into actionable data narrative practices. Our objective is to direct the reader towards research papers that offer practical and actionable contributions for the implementation of the concepts in the data narrative model in an automated manner. We aim to provide guidance to readers who are interested in implementing the model in a real-world context, by highlighting papers that offer tangible solutions and strategies for automating the concepts of the data narrative.

1.2.1 Factual layer

Exploration. The literature about how data workers explore and analyze data is abundant. This activity, sometimes called discovery-driven analysis [144], interactive data analysis [118] or interactive data exploration [50], is characterized by the building of ad-hoc exploration actions (queries, model construction, etc.), conceived incrementally in a hypothesis-driven fashion, to identify regions of anomalies also interesting data in large amounts of data. It is often described as a tedious and difficult task, especially for novice users. Several papers addressed automatic exploratory data analysis and OLAP query enrichment. The works that instantiate explorations in data narration are listed below.

From the exploratory data analysis side, Chanson et al. [31] developed an automated algorithm that assists an analyst in exploring a data set by suggesting the most interesting and coherent set of queries that are estimated to be completed under a time constraint. Milo et al. [119] reviewed recent work for automating exploratory data analysis and then identified three main categories for recommended systems for suggesting specific and high-utility exploratory operations. Firstly, a data-driven that uses heuristic notions of interestingness and employs them to find data subsets conveying interesting patterns, data visualizations, or data summaries. Secondly, log-based systems that leverage a log of former exploratory operations, performed by the same or different users, in order to generate more personalized exploratory data analysis recommendations. Thirdly, hybrid and the data set being currently explored.

In addition to recommender systems, several papers proposed fully automating the exploration process by leveraging advances in deep neural networks. For example, Dibia et al. [45] presented a system for auto-generating data visualizations based on a sequence-to-sequence recurrent neural network model. Milo et al. [51,117] suggest generating entire EDA sessions by formulating EDA as a control problem, and solving it using deep reinforcement learning.

Regarding data analysts practices, Wongsuphasawat et al. [173] conducted semi-structured interviews with 18 analysts from academic and industrial settings to understand exploratory data analysis practices. They observed that all participants engage in profiling (understanding what the data contain and assessing data quality) across all of their analyses, while discovery (gaining new findings) only reliably occurs in open-ended analyses, which participants perform less often.

Collector. Several works addressed the implementation of OLAP operators. For instance, Sarawagi et al. [145] proposed three advanced operators to automate data analysis tasks that are currently handled through manual exploration: (i) Diff operator automates the drill-down operation to find causes for drops or increases observed at an aggregated level, (ii) Relax operator automates roll up operation to view the problem case in context of combinations of other dimensions using a succession of selectionand, (iii) Surprise operator helps a user to quickly familiarize himself with the significant features of a OLAP data cube in a manner that adapts to the prior context of the user.

Furthermore, Kraiem et al. [93] proposed new OLAP operators that enhance existing solutions for OLAP analysis involving a reflexive relationship on the fact instances and dealing with missing values on dimension members: Null-Drilldown, Null-Rollup and Null-Select handle large volumes of data among which a great amount of missing data is involved, FDrilldown and FRollup provide solutions for handling an intuitive navigation between different levels within the fact.

Three operators are proposed by ATENA [52], a system that takes an input dataset and auto-generates a compelling exploratory session, presented in an exploratory data analysis notebook. Two operators encapsulate the most common SQL queries like FILTER to select data tuples that match a criteria and GROUP to aggregate the data. The last operator is BACK that allows the agent to backtrack to the previous dis- play in order to take an alternative exploration path.

Gkesoulis et al. [68] demonstrated how to enrich query answering of an OLAP query. The first assessment of the current state of affairs refers to the execution of the original query then, they exploit the selection conditions of the original query and automatically generate complementary queries that compare its results with the results of queries having similar values. Finally, they drill in the grouping levels of the original result to see the breakdown of its (aggregate) measures and understand its internal structure to provide further analysis of the results.

Among the plethora of possibilities, choosing the right collector can be done through automated learning, like in the case of model construction [56], various recommendation techniques [5, 109, 118], or be built-in in the user's specification [163].

Finding. To retrive findings from data, Wang et al. [168] assign a *score* to determine the importance of the data extracted based on two factors: significance and impact. Zgraggen et al. [179] categorized findings in five classes: shape, mean, variance, correlation, and ranking. Each class has its own hypothesis testing scheme for finding validation.

We discovered a variety of methods for extracting findings from the data. First, Sarawagi et al. [144] outlined exceptions for obtaining findings from a piece of information's degree of surprise. The surprise value of information is therefore a combination of the following three values: (i) SelfExp represents the information's surprise value relative to others at the same aggregation level, (ii) InExp represents the degree of surprise somewhere beneath this information if we zoom in to more detailed hierarchies from the entity, and (iii) PathExp represents the degree of surprise for each path of zooming in to more detailed hierarchies from the information. Second, Marcel et al. [107] defined a degree attributed to a piece of information regarding the curiosity and surprise it generates. This degree is characterized by four aspects: (i) relevance as a measure for the user's curiosity based on area of interest and user's intent, (ii) novelty to annotate if the

information is visited or not, (iii) surprise to discover new values and, (iv) peculiarity to distinguish information to others. Thirdly, Deutch et al. [44] proposed an ExplainED system that analyzed a set of data in order to detect what elements thereof are particularly interesting, and produced a corresponding textual explanation by evaluating the interestingness of the given view using several measures capturing different interestingness facets, then computing the Shapely values of the elements in the view, w.r.t. the interestingness measure yielding the highest score.

1.2.2 Intentional layer

Analytical Question. The intention of the data narrator during data exploration, namely the formulation of an analytical question, is not effectively addressed. With the exception of high-level intentions identified in [161, 163] that contribute to implementing an analytical question, these intentions are formally described in the form of operators of a high-level language, and some of them are accompanied by concrete implementation: (i) describe [33] when the user's intention is to learn more about a set of data, (ii) predict when the user's intention is to predict future values, (iii) explain when user aims at understanding the cause of a phenomenon and, (iv) assess [59, 152] when the user's intention is to compare, add more data to an ongoing analysis and, (v) suggest when user aims at asking for guidance.

Message. The crafting of messages, i.e., the production of easily understandable interesting or important information extracted through data analysis [35], is rarely addressed. There are 3 categories of works, concerning how messages are handled: (i) works limiting to the selection of a subset of findings [35, 99, 149, 150, 168], (ii) works focusing on common patterns and exceptions [102] and, (iii) works focusing on measurement assimilation towards crafting messages [81].

Different narrative patterns identified in [9] help the data narrator choose the way of message communication depending on his intention. For example, if the intent of the data narrator is to *persuade* and *convince* audience, he can use one of the following patterns: compare, concretize, and repetition. In addition to the convince pattern, they proposed additional patterns for communicating messages: reveal, repetition, slowing down and speeding-up to structure the *sequencing of messages* and arguments; familiar setting, make-a-guess, defamiliarization, convention breaking, silent data, physical metaphors to *integrate the audience* to the story; gradual reveal, slowing down, speed-up, concretize, breaking-the-fourth-wall, humans-behind-the-dots, rhetorical question, call-to-action, familiarize pay attention to *perceive and reflect* on the message; and rhetorical question, call-to-action, breaking-the-4th-wall, make-a-guess, exploration to *involve reader* to be a part of the story.

1.2.3 Structural layer

Plot. The data excerpts gathered from the analysis phase must be assembled into a plot: a sequence well organized. Ordering, forming logical relations, designing flow, formulating a message, and creating the dénouement are all part of the generation of the plot of data narratives.These tasks, which are often interconnected, may be completed sequentially,

concurrently, or in multiple iterations [98]. We found that different structures and organizations are possible to convey information to the audience in an understandable way because narratives, like sentences, have their own internal structure.

Several works, under various names, deal with structuring the plot of narrative [68, 149,150,168]. For instance, Gkesoulis et al. [68] generated a rigid plot with three acts: put in context, build up protagonist's action and conclude a resolution. Datashot [168] choose messages that have the same topic, then selects messages based on density-based top-n algorithm balancing message diversity and importance. Calliope system [150] generated the plot of the narrative using the logic-oriented Monte Carlo tree search algorithm that explores the data space to generate a series of messages in logical context to build a story. Autoclips system [149] generated a data video by reorganizing the sub-sequences within the parallel structure (repeated pattern to tell a story) and juxtaposing two visualizations supported by the same type of messages in one scene. The choice of transitions between two scenes is based on visualization transitions proposed in [80].

Several works addressed the sequencing and ordering of narrative points to structure a plot. For example, Hullman et al. [80] proposed a graph-driven approach to find effective sequences for narrative visualizations. Brehmer et al. [24] analyzed the design space for storytelling timelines i.e., ways to convey multiple narrative points, explained and presented considerations for authoring cohesive stories containing a variety of narrative points including transitioning between different timeline designs.

Act. Gkesoulis et al. [68] generated three acts for each data story: the first providing contextualization for the characters as well as the incident that sets the story on the move, the second where the protagonists and the rest of the roles build up their actions and reactions and the third where the resolution of the film is taking place. In more details, the first assessment of the current state of affairs refers to the execution of the original query provided by the user, then, they exploit the selection conditions of the original query and automatically generate complementary queries that compare its results with the results of queries having similar values to provide contextualization and finally, they drill in the grouping levels of the original result to see the breakdown of its (aggregate) measures and understand its internal structure to provide further analysis of the results.

Episode. Bach [9] introduces so-called narrative design patterns, where a narrative pattern is defined by "a narrative pattern is a low-level narrative device that serves a specific intent. A pattern can be used individually or in combination with others to give form to a story." Five major patterns of group are identified: argumentation, flow, framing, emotion, engagement. Importantly, these patterns are not specifically related to a visualization or interaction medium. Besides the pattern to generate an episode, Gkesoulis et al. [68] organized the data narrative as movies in terms of acts, with each act including several episodes all serving the same purpose. Each episode comes with queries in its background and holds all the important information (highlights, visual, and audio parts). Blount et al. [19] aim to identify patterns used for data narrative by examining 67 stories from both professional journalists including award-winning data stories and data-science aware students. The patterns they look for are those introduced in [9]. Overall, they note that a large majority of patterns used are from the argumentation and flow categories.

1.2.4 Presentational layer

Visual narrative. There are many ways to visualize a narrative. This recently opened a new branch of research suggesting that visual narrative comprehension involves many cognitive tasks, among them visual perception, attention, and language [39]. Following this direction, several works have been introduced to provide actionable visual narratives in different application domains. In [168], the authors categorize the research on visual narratives in two branches, one about authoring systems that ease the design process of dashboards for users having an understanding of data and how to present them (e.g., [24, 166]), and the other about automatic deploy of data narratives.

Besides the many approaches addressing the design nuances of narratives within disparate domains (e.g., spatio-temporal trajectories [177], weather conditions [139], social networks [115], healthcare [18], etc.), research effort has been spent on representing narratives with linear, concurrent, or overlapping storylines. In [80], the authors analyze how sequencing choices affect narrative visualization and introduce an algorithmic approach to visualization sequence support. The approach specifies visualization states as nodes in a graph to compare nodes and evaluate potential transition. The authors identify an objective function based on the principle of maintaining consistency between visualizations and apply weights to edges (transitions) in the graph to allow assessment of the quality of transitions at the local level. Indeed, structures that are simpler (i.e., contain few, more homogeneous groupings) and more consistent (e.g., parallel transitions within groupings) are usually preferred by users [82]. In [89], the authors explore nonlinear storytelling to portray events of a story out of chronological order. While this narrative style is acclaimed in movies, communicating nonlinear narratives is difficult. The authors introduce two techniques: story curves to visualize nonlinear events against their actual chronological order and story explorer to complement storytelling with contextual information.

Since many visual narratives have been conceived, we also refer the reader to comprehensive studies [147, 155].

Dashboard. Dashboards are widely used in data-intensive applications such as business intelligence, operation monitoring, and urban planning [103]. However, building dashboards requires expertise both in the application and visualization domains. The dashboard design can be customizable, model driven, and adaptive [164]. In doing so, authoring tools (i.e., programs or environments designed for the development of dashboard; e.g., Wirecloud[8]) help users in composing their dashboard. Noticeably, recent research direction involves the extraction of visual dashboard out of human sketches [103] through deep learning models that infer and help users conceptualizing their design intention. Model-driven generation is based on data about users, tasks, or goals. For instance, in [90, 131], the authors take as input models of the business process and KPIs to describe and generate a dashboard. To describe query results, in [60], the authors organize a visual dashboard as a composition of three areas, one about data, one about insights, and one about user-friendly charts adapted to the data at hand. In [59], the authors introduce a *notebook-like* interface to compare query results in form of small multiples. Adaptive design [14] changes dashboard based on environmental changes. We refer the reader to [164]

[8] https://www.fiware.org/events/creating-advanced-dashboards-using-wirecloud/

for a systematic survey on tailoring dashboards to data and user requirements.

Recently, dashboards [8,12] have witnessed an increasing interest in the field of medical data.

Dashboard component. A comprehensive description depicted in [73] details the process of choosing the right data visualization: bubble chart, grouped column graphs are perfect for comparing one or many value sets, bubble graph, dendrogram and tree-map for representing different groups that share some form of similarity, pie chart, grouped column graph, tree-map, and dendrogram can be used for showing the part-to-whole composition, the grouped column graph can be used to highlight order, bubble graph and dendrogram can be used to understand the relationship between value sets, heat map, single-line graph and marked line graph for looking at how data is distributed, heat-map to show how geographical areas of a map compare to one another based on a given criterion.

In [60], the authors provide guidelines and an algorithm to pick the proper dashboard component depending on the dimensionality of the result, the number of numerical and categorical attributes, and the cardinality of each attribute. The authors adopt heuristics to decide whether to use or not each chart type for a given query result, where a score is assigned to each chart type depending on the features of the dataset to be visualized. For instance, bubble charts are considered to be suitable to visualize n-dimensional data if the bubble size is mapped to a numerical attribute and the bubble color is mapped to either a numerical attribute or a categorical attribute. Following these guidelines, [59] leverage small multiples—a series of similar charts using the same scale and axes—to enable a straightforward comparison of different facets of a dataset.

1.3 Reverse engineering data narratives

A good analysis support for a data narrative can come in the form of a Python notebook because of its reproducible results, documentation abilities, visualization possibilities, collaborative features, and interoperability with version control systems. It allows analysts to combine code, data, and explanations seamlessly to create a narrative around the data. A simple example for a data narrative[9] about *the COVID pandemic in a French region* documented by a data journalist in the form of a notebook[10]. The structure of a data narrative typically comprises elements such as titles, textual content, and visualizations, while a notebook is structured around cells containing both text and code.

Reverse engineering a data narrative involves deconstructing the narrative to identify the underlying data, analysis goal, messages and the concepts of the data narrative. We believe that the result of reversing a data narrative can be represented in a data structure, where the concept serves as the key and the corresponding data structure serves as the value. Algorithm proposed for reverse engineer a data narrative is depicted in 1. Elaboration on the algorithm's functions is provided in the sections that follow its presentation. In this section, we detail how a data narrative can be reverse-engineered from its support analysis. The algorithm takes as an input a visual narrative and a notebook python containing the analysis and returns a map that maps two concepts.

[9] https://tinyurl.com/24ubaanu(in french)
[10] https://tinyurl.com/yc5chu57(in french)

It's important to emphasize that the core functions detailed in this algorithm are presented in a simplified manner. Nevertheless, it's noteworthy to mention that there exist more advanced implementations beyond the scope of what has been described here.

Steps. To reverse engineer a data narrative, it is important to identify the components of the visual narrative. We considered data narrative as a visual narrative, where a single dashboard includes various dashboard components, including textual elements and graphical figures. Then we moved to find the primary concept of the data narrative the "messages" that are being conveyed in the data narrative. Subsequently, we proceeded to identify the structural concepts inherent to the data narrative. Moving to the data narrative support analysis, we identified the factual concepts. In the final step, an alignment between the messages and the concepts they related to was emerged, including the findings, collectors, analytical questions and visualizations.

Identifying presentational concepts. (see lines 1-8 in Algorithm 1)

We considered data narrative as a visual narrative, where a single dashboard includes various dashboard components, including textual elements and graphical figures.

Identifying intentional concepts. (see lines 9-13 in Algorithm 1)

One possible method to identify and find the message is to analyze each sentence narrated in the data narrative. A sentence that starts with a capital letter and ends with a punctuation mark, such as a period or question mark, is considered as a potential message. Once the main message is identified, the next step is to identify the analysis goal and the analytical questions. To accomplish this task, one could examine the analysis support of the data narrative. For instance, identifying synonyms and related terms for the term "goal" or "objective" and searching for them within the text is a viable approach to locating the goal in the documented process. Similarly, identifying a pattern for analytical questions, such as those beginning with "what", "when", or "does" is necessary to identify such questions within the documented process. Note that when a concept and its value are identified, they are added to the map without explicitly mentioning them.

Identifying structural concepts. (see lines 14-19 in Algorithm 1)

Our approach to identify the act in the data narrative involves treating each section as an act. Specifically, we define each act as the content that falls between two subtitles. Once we have identified each act in the data narrative, we then determine the associated question for each act. To do this, we match the characters in each analytical question to the characters in each section of the data narrative. If a section and an analytical question share a significant number of common characters, then they are considered associated. After identifying the acts and associated questions in the data narrative, the next step is to identify the collector and its related question for each act. This can be achieved by analyzing the characters of the analytical question and the components of the collector.

Identifying factual concepts. (see lines 20-27 in Algorithm 1)

Similar to the step of identifying the associated question for each act, we match the characters of the analytical question to the collector components to determine the collector and related question for each act. Collector component means the elements of a query as the name of function, the name of columns, name of variables or the dimension and measure in multidimensional spaces, etc. Through the comparison of these components with the characters in the analytical questions, we discover a significant overlap between the collectors and analytical questions. This overlap allows us to identify which collectors are used to implement the analytical question. In our work, we consider each cell code as

a collector, and each text cell that appears after the cell code as a finding.

Connecting message to related concepts. (see lines 28-42 in Algorithm 1)

To establish a connection between a message and its associated concepts, we initially identify the "act" to which the message belongs. This is accomplished by locating the message within the corresponding act. Subsequently, we proceed to determine the pertinent questions and their respective collector and finding related to that act. For each finding, we test whether the message is composed of that finding or not, by comparing the characters and measures in the message to the words in the finding. If there is no finding that matches the message, we check if the visualization handle the characters and measures of the message. If the visualization is capable of doing so, it will be considered as a finding. It's important to note that we considered that the plot of the data narrative is conveyed through the visual narrative, where each act is presented within a dashboard, and each episode is depicted within a dashboard component. The outcome of the reverse engineering algorithm yields two distinct maps. The first map encompasses all the concepts featured in the data narrative along with their corresponding values. Meanwhile, the second map records associations between pairs of concepts. For instance, it may link a collector to its corresponding finding, storing this relationship as a value within the map.

Algorithm 1 Reverse engineer a data narrative from its analysis support

Require: A visual narrative D and a notebook P

Ensure: A Map map that finds the concepts and their values in P, an associative assoc_map that maps two concepts.

```
// Identifying presentational concepts
Visual narrative = D
for section ∈ D do
    Dashboard = section
    for element ∈ section do
        Dashboardcomponent = element
    end for
end for
// Identifying intentional concepts
g findGoal(P)
AQ[] aqs = findAQ(P)
map.add(Goal, g)
map.add(Analytical question, aqs)
// Identifying structural concepts
Act[] acts = findActs(P)
for act ∈ acts do
    AQact = findAQAct(act)
    assoc_map.add(act, AQact)
end for
// Identifying factual concepts
Collector[] collectors = findCollectors(P)
for collector ∈ collectors do
    AQ AQcollector = findAQcollector(collectors, aq)
    assoc_map.add(AQcollector, collector)
    Finding finding = findFindings(collector)
    assoc_map.add(Collector, findings)
end for
// Connecting message to related concepts
for message ∈ messages do
    act = findMessageinAct(message)
    AQconcernAct = assoc_map.get(act)
    collectorAQ = assoc_map.get(AQconcernAct)
    Findingcollector = map.get(collectorAQ )
    if finding is NULL then
        for Visualization ∈ P do
            if Message ∈ Visualization then
                assoc_map.add (Message,Visualization)
            end if
        end for
    end if
end for
return map, assoc_map
```

```
1: function FindGoal(P)
2:     // Objective: a sentence containing a synonym of "objective"
3:     synonyms ← set of synonyms for "objective"
4:     for s ∈ P do
5:         for w ∈ s  do
6:             if w is a synonym of any word in synonyms then
7:                 return s
8:             end if
9:         end for
10:    end for
11:    return empty string
12: end function
```

```
1: function FindAQ(P)
2:     Patterns ← Does, How, what is, etc
3:     questions ← empty list
4:     for each sentence s in P do
5:         for each pattern p in Patterns do
6:             if s matches p then
7:                 questions.add(s)
8:             end if
9:         end for
10:    end for
11:    return questions
12: end function
```

```
1: function FindMessages(D)
2:     messages ← empty list
3:     for each sentence s in D do
4:         if s starts with a letter and ends with a period then
5:             messages.add(s)
6:         end if
7:     end for
8:     return messages
9: end function
```

```
1: function FINDACTS(D)
2:     acts ← empty array
3:     subtitles ← all subtitles in P
4:     for i ← 1 to length(subtitles) − 1 do
5:         start ← end index of subtitle i
6:         end ← start index of subtitle i + 1
7:         act ← text between start and end in P with content
8:         acts.add(act)
9:     end for
10:    return acts
11: end function
```

```
1: function FINDAQSECTION(act)
2:     aq ← empty array
3:     analyticalQuestions ← all analytical questions
4:     for question ∈ analyticalQuestions do
5:         if act and question share a significant number of common characters then
6:             append question to aq
7:         end if
8:     end for
9:     return aq
10: end function
```

```
1: function FINDCOLLECTOR(P)
2:     collectors ← empty array
3:     for cell in all cell codes in P do
4:         collectors.add(cell)
5:     end for
6:     return collector
7: end function
```

```
1: function FINDAQCOLLECTOR(collectors, aq)
2:     aq_collector ← empty list
3:     for collector in collectors do
4:         for component in collector do
5:             if component contains significant characters of aq then
6:                 aq_collector.add(collector)
7:             end if
8:         end for
9:     end for
10:    return aq_collector
11: end function
```

```
1: function FindFinding(collector)
2:     Findings ← empty array
3:     Locate collector in P
4:     for text_cell ← all text cells after the cell code do
5:         findings.add(cell)
6:     end for
7:     return findings
8: end function
```

```
1: function FindMessageInAct(message, acts)
2:     for act in acts do
3:         if message in act then
4:             return act
5:         end if
6:     end for
7:     return null
8: end function
```

```
1: function FindVizMessage(message, viz)
2:     characters ← FindCharacters(message)
3:     measures ← FindMeasures(message)
4:     canHandle ← true
5:     for character in characters do
6:         if character not in viz.supportedCharacters then
7:             canHandle ← false
8:         end if
9:     end for
10:    for measure in measures do
11:        if measure not in viz.supportedMeasures then
12:            canHandle ← false
13:        end if
14:    end for
15:    if canHandle then
16:        return viz as finding
17:    else
18:        return null
19:    end if
20: end function
```

```
1: function FindCharacters(message)
2:     characters ← empty list
3:     for word in message.words do
4:         if word is a noun then
5:             characters.append(word)
6:         end if
7:     end for
8:     return characters
9: end function
```

```
1: function FindMeasures(message)
2:     measures ← empty list
3:     for word in message.words do
4:         if word is a numerical value then
5:             measures.append(word)
6:         end if
7:     end for
8:     return measures
9: end function
```

1.4 Usage of the web application

We present a functional description for supporting data narrative generation with a web application containing a set of text fields titled to understand the story generation path, and a *log and console* to keep track of narrative details. To start crafting a story, the user (or data narrator) clicks *start new story* button, defines the story goal by filling up *analysis goal* and clicking on *define analysis goal* to log the goal into the story's logs. They then pose an analytical question and click on *add new analytical question* to log the question. To answer the analytical question, the user tries different collectors to fetch the data stored in a database by choosing either *create SQL query* or *create describe collector*. The user writes the collector's query and gets the result as a set of tuples and simple charts by clicking on *evaluating this query*. They look over the facts retrieved by the collector, choose the important findings to turn into a message by clicking *validating collector finding*. Important findings are copied into the *message* text area to allow the data narrator to edit it, before logging it. For each message, the user is responsible to fill up its *measure(s)* and *character(s)*. These measures and characters can be recalled later while writing new episodes. When a message is created, the user is allowed to organize the story structure by creating different acts and episodes. The user can create, add and attach different episodes to a specific act, while each episode narrates only one message. The manner of creating and organizing acts and episodes is left to data narrator. At the end, the user can download the story as a PDF document by clicking on *PDF of narrative*. Also, a *notebook SQL* can be generated to document the SQL data exploration.

1.5 Examples for Highlights

Additional concrete examples illustrating the concept of Holistic Highlights and Elementary Highlights are instantiated and displayed in Table A.8 and Table A.9. These examples further exemplify the distinction between the two types of highlights and provide visual representations of how they manifest in real-world data scenarios.

The **Highlight Type**	for **Indicator**	tested via **Algo**	*{Supporting text*	*Supporting role}**	fits under the model **Resulting Model**	with **Score Label =**	**Score Value**	with elementary hl's	Comments
Distribution of values	`SumSales`	Shapiro Wilk			*Normal*	p-value	0.001		Similarly for other distributions, e.g., Power Law, Uniform, ... with the respective test
Correlation	`SumSales`	Spearman	wrt measure	`cityPopulation`	*Significant*	r	0.76		
Trend, Seasonality	`SumSales`	STL	over attribute	`month`	*False*			If true, the sumSales is split in *<Sales.Tr, Sales.S, Noise>*	
Regression formula	`SumSales`	linRegr	over attribute	`cityPopulation, month`	*S=f(p,t)*	MSE	9.6		Requires extending the result with city.population. There exists a formula for explaining a measure.
Uni(bi) modality	`SumSales`	Uni(bi) ModCheck	over attribute	`month`	*True* (uni/bi mod. Exists)			Point of peak(s)	
Top-k values	`SumSales`	topK check			*True (i.e., top-k values do exist)*			Top-k points	Similarly for bottom-k
Peer-domination	`SumSales`	Peer Dom Check	over attributes	`city, month`	*True (i.e., peerDominator values do exist)*			dominant characters	
Total sum Mega-Contribution	`SumSales`	MegaContrib	over attributes	`city, month`	*True (i.e., mega contributor values do exist)*			Mega-contributor fact	
Surprise	`SumSales`	Extrapolate 3 Past Years	over attributes	`city, month`	True (i.e., several facts deviate expected value)	MSE	4.8	Surprising facts	Several facts with significant deviations from expected values

Figure A.8: Examples of holistic highlights

The **Dimensional Value**	*over* **Dimensional Attribute**	With a **Indicator Value**	For **Indicator Type**	serves as the **Highlight Type**	with a **Score Type**	having **Score**	Comments
<April 20022>	Month	68	SumSales	1st peak of unimodality	*Peak Rank*	1	You need a single-grouper query, such that Measure = f(Time)
<June 2022>	Month	62	SumSales	2nd peak of bimodality	*Peak Rank*	2	You need a single-grouper query, such that Measure = f(Time)
<Athens, June 2022>	City, Month	54	SumSales	Top-k point	*Rank*	2	
<Athens>	City	54	SumSales	peer-dominator over its sibling characters	*Pct of dominated peers*	100%	For each month For each city c' != Athens Athens dominates c'
<June 2022>	Month	32	SumSales	peer-dominator over its sibling characters	*Pct of dominated peers*	80%	For each city For 80% of months m' != June 2022 June 2022 dominates m'
<June 2022, Athens>	City, Month	12	SumSales	a mega-contributor of the total sum	*Pct of total sum*	25%	
<Athens>	City	2100	SumSales	a mega-contributor of the marginal sum of cities	*Pct of total sum*	75%	Fix a column col. For the marginal sum of all cells of this col, a single cell is a mega contributor. So, if @the holistic level you say the col is fixed, there is a single character to determine the h/l; else: 2 of them.
<April 2022, Rhodes>	City, Month	10	SumSales	Surprising cell	*Abs Diff Ratio*	28%	
				time-series change point			You need a single-grouper query, such that Measure = f(Time)

Figure A.9: Examples of elementary highlights

Bibliography

[1] Serge Abiteboul, Omar Benjelloun, Tova Milo, and Victor Vianu. *Theoretical Foundations of Data Integration.* Morgan & Claypool Publishers, 2009.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[3] Andrea Lezcano Airaldi, Emanuel Irrazábal, and Jorge Andrés Diaz-Pace. Narrative visualizations best practices and evaluation: A systematic mapping study, 2022.

[4] Ergun Akleman, Stefano Franchi, Devkan Kaleci, Laura Mandell, Takashi Yamauchi, and Derya Akleman. A theoretical framework to represent narrative structures for visual storytelling. In *Bridges*, pages 129–136, 2015.

[5] Julien Aligon, Enrico Gallinucci, Matteo Golfarelli, Patrick Marcel, and Stefano Rizzi. A collaborative filtering approach for recommending OLAP sessions. *Decis. Support Syst.*, 69:20–30, 2015.

[6] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Trans. Vis. Comput. Graph.*, 25(1):22–31, 2019.

[7] Robert A. Amar, James Eagan, and John T. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA*, pages 111–117. IEEE Computer Society, 2005.

[8] Stelios Andreadis, Gerasimos Antzoulatos, Thanassis Mavropoulos, Panagiotis Giannakeris, Grigoris Tzionis, Nick Pantelidis, Konstantinos Ioannidis, Anastasios Karakostas, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. A social media analytics platform visualising the spread of COVID-19 in italy via exploitation of automatically geotagged tweets. *Online Soc. Networks Media*, 23:100134, 2021.

[9] Benjamin Bach, Moritz Stefaner, Jeremy Boy, Steven Drucker, Lyn Bartram, Jo Wood, Paolo Ciuccarelli, Yuri Engelhardt, Ulrike Koppen, and Barbara Tversky. *Narrative Design Patterns for Data-Driven Storytelling.* 2018.

[10] Benjamin Bach, Zezhong Wang, Matteo Farinella, Dave Murray-Rust, and Nathalie Henry Riche. Design patterns for data comics. In *CHI*, 2018.

[11] Mieke Bal. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press, Toronto, second edition, 1997.

[12] Erica Barbazza, Damir Ivanković, Sophie Wang, Kendall Jamieson Gilmore, Mircha Poldrugovac, Claire Willmington, Nicolas Larrain, Véronique Bos, Sara Allin, Niek Klazinga, et al. Exploring changes to the actionability of covid-19 dashboards over the course of 2020 in the canadian context: Descriptive assessment and expert appraisal study. *Journal of medical Internet research*, 2021.

[13] Leilani Battle and Jeffrey Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Comput. Graph. Forum*, 38(3):145–159, 2019.

[14] Orlando Belo, Paulo Rodrigues, Rui Barros, and Helena Correia. Restructuring dynamically analytical dashboards based on usage profiles. In *ISMIS*, volume 8502 of *Lecture Notes in Computer Science*, pages 445–455. Springer, 2014.

[15] Laure Berti-Équille and Mouhamadou Lamine Ba. Veracity of big data: Challenges of cross-modal truth discovery. *ACM J. Data Inf. Qual.*, 7(3):12:1–12:3, 2016.

[16] Bruno Bettelheim. *The Uses of Enchantment: The Meaning and Importance of Fairy Tales*. Thames & Hudson, 1976.

[17] Tijl De Bie. Subjective interestingness in exploratory data mining. In *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013. Proceedings*, volume 8207 of *Lecture Notes in Computer Science*, pages 19–31. Springer, 2013.

[18] Sacha E Bleeker, Gerarda Derksen-Lubsen, Astrid M van Ginneken, Johan Van Der Lei, and Henriëtte A Moll. Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics. *BMC medical informatics and decision making*, 6(1):1–7, 2006.

[19] Tom Blount, Laura Koesten, Yuchen Zhao, and Elena Simperl. Understanding the use of narrative patterns by novice data storytellers. In *CHIRA*, pages 128–138. SCITEPRESS, 2020.

[20] Raphaël Bonaque, Tien Duc Cao, Bogdan Cautis, François Goasdoué, J. Letelier, Ioana Manolescu, O. Mendoza, S. Ribeiro, Xavier Tannier, and Michaël Thomazo. Mixed-instance querying: a lightweight integration architecture for data journalism, 2016.

[21] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[22] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Davide Mazza. Exploratory search framework for web data sources. *VLDB J.*, 22(5):641–663, 2013.

[23] Richard Brath and Michael Peters. Dashboard design: Why design is important. *DM Direct*, 2004.

[24] Matthew Brehmer, Bongshin Lee, Benjamin Bach, Nathalie Henry Riche, and Tamara Munzner. Timelines revisited: A design space and considerations for expressive storytelling. *IEEE Trans. Vis. Comput. Graph.*, 23(9):2151–2164, 2017.

[25] Andrei Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[26] Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. New Riders, 2013.

[27] Daniel Calegari. Computational narratives using model-driven engineering. In *XLVIII Latin American Computer Conference, CLEI 2022, Armenia, Colombia, October 17-21, 2022*, pages 1–9. IEEE, 2022.

[28] Sheelagh Carpendale, Nicholas Diakopoulos, Nathalie Henry Riche, and Christophe Hurter. Data-driven storytelling (dagstuhl seminar 16061). *Dagstuhl Reports*, 6(2):1–27, 2016.

[29] Sylvie Cazalens, Julien Leblay, Ioana Manolescu, Philippe Lamarre, and Xavier Tannier. Computational fact-checking: a content management perspective. *Proc. VLDB Endow.*, 11(12):2110–2113, 2018.

[30] Marie Chagnoux. La datavisualisation, double point d'entrée du datajournalisme dans la PQR (in french). *Interfaces numériques*, 9(3), 2020.

[31] Alexandre Chanson, Ben Crulis, Nicolas Labroche, Patrick Marcel, Verónika Peralta, Stefano Rizzi, and Panos Vassiliadis. The traveling analyst problem: Definition and preliminary study. In *DOLAP*, volume 2572 of *CEUR Workshop Proceedings*, pages 94–98. CEUR-WS.org, 2020.

[32] S.B. Chatman. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell paperbacks. Cornell University Press, 1980.

[33] Antoine Chédin, Matteo Francia, Patrick Marcel, Verónika Peralta, and Stefano Rizzi. The tell-tale cube. In *ADBIS*, volume 12245 of *Lecture Notes in Computer Science*, pages 204–218. Springer, 2020.

[34] Qing Chen, Shixiong Cao, Jiazhe Wang, and Nan Cao. How does automation shape the process of narrative visualization: A survey on tools. *CoRR*, abs/2206.12118, 2022.

[35] Siming Chen, Jie Li, Gennady L. Andrienko, Natalia V. Andrienko, Yun Wang, Phong H. Nguyen, and Cagatay Turkay. Supporting story synthesis: Bridging the gap between visual analytics and storytelling. *IEEE Trans. Vis. Comput. Graph.*, 26(7), 2020.

[36] Zheng Chen, Fan Lin, Huan Liu, Wei-Ying Ma, and Liu Wenyin. User intention modelling in web applications using data mining. *World Wide Web*, 5, 2002.

[37] Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. Is everything in order? A simple way to order sentences. In *EMNLP*. Association for Computational Linguistics, 2021.

[38] Antoine Chédin, Matteo Francia, Patrick Marcel, Verónika Peralta, and Stefano Rizzi. The tell-tale cube. In *ADBIS*, 2020.

[39] Neil Cohn and Joseph P. Magliano. Editors' introduction and review: Visual narrative research: An emerging field in cognitive science. *Top. Cogn. Sci.*, 12(1):197–223, 2020.

[40] Yiming Cui, Wanxiang Che, Wei-Nan Zhang, Ting Liu, Shijin Wang, and Guoping Hu. Discriminative sentence modeling for story ending prediction. In *AAAI*, pages 7602–7609. AAAI Press, 2020.

[41] Raphaël da Silva. Mortalité- totaux, moyenne et segmentation avec pandas.ipynb. https://tinyurl.com/yc5chu57, 2020. Online; accessed 20 January 2021.

[42] Raphaël da Silva. Toutes causes confondues, la covid a tué jusqu'à cinq fois plus d'alsaciens pendant la crise. https://tinyurl.com/24ubaanu, 2020. Online; accessed 20 January 2021.

[43] Evert de Haan, Peter Verhoef, and Thorsten Wiesel. The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing*, 32, 2015.

[44] Daniel Deutch, Amir Gilad, Tova Milo, and Amit Somech. Explained: Explanations for EDA notebooks. *Proc. VLDB Endow.*, 13(12):2917–2920, 2020.

[45] Victor Dibia and Çagatay Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE Computer Graphics and Applications*, 39(5):33–46, 2019.

[46] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support. *IEEE Transactions on Visualization and Computer Graphics*, 24, 2018.

[47] Apiwan Duangphummet and Puripant Ruchikachorn. Visual data story protocol: Internal communications from domain expertise to narrative visualization implementation. In *VISIGRAPP*, pages 240–247. SCITEPRESS, 2021.

[48] Vanessa Echeverría, Roberto Mart'inez Maldonado, and Simon Buckingham Shum. Towards data storytelling to support teaching and learning. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. ACM, 2017.

[49] Cameron Edmond and Tomasz Bednarz. Three trajectories for narrative visualisation. *Vis. Informatics*, 5(2):26–40, 2021.

[50] Philipp Eichmann, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. Idebench: A benchmark for interactive data exploration. In *SIGMOD*, pages 1555–1569. ACM, 2020.

[51] Ori Bar El, Tova Milo, and Amit Somech. ATENA: an autonomous system for data exploration based on deep reinforcement learning. In *CIKM*, pages 2873–2876. ACM, 2019.

[52] Ori Bar El, Tova Milo, and Amit Somech. Automatically generating data exploration sessions using deep reinforcement learning. In *SIGMOD Conference 2020, online conference [Portland, OR, USA]*, pages 1527–1537. ACM, 2020.

[53] David K Elson. *Modeling narrative discourse*. PhD thesis, Columbia University, 2012.

[54] EL Outa Faten, Marcel Patrick, Peralta Veronika, and Vassiliadis Panos. Highlighting the importance of intentional aspects in data narrative crafting processes. *Information Systems Frontiers*, pages 1–17, 2023.

[55] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. Aurum: A data discovery system. In *ICDE*, pages 1001–1012. IEEE Computer Society, 2018.

[56] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pages 2962–2970, 2015.

[57] Stephen Few. *Information dashboard design: The effective visual communication of data*. O'reilly Sebastopol

[58] Matteo Francia, Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Insight-based vocalization of OLAP sessions. In *ADBIS*, volume 13389 of *Lecture Notes in Computer Science*, pages 193–206. Springer, 2022.

[59] Matteo Francia, Matteo Golfarelli, Patrick Marcel, Stefano Rizzi, and Panos Vassiliadis. Assess queries for interactive analysis of data cubes. In *EDBT*, pages 121–132. OpenProceedings.org, 2021.

[60] Matteo Francia, Patrick Marcel, Verónika Peralta, and Stefano Rizzi. Enhancing cubes with models to describe multidimensional data. *Information Systems Frontiers*, 2021.

[61] Matteo Francia, Patrick Marcel, Verónika Peralta, and Stefano Rizzi. Enhancing cubes with models to describe multidimensional data. *Inf. Syst. Frontiers*, 24(1):31–48, 2022.

[62] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning semantic hierarchies via word embeddings. In *ACL*, pages 1199–1209. The Association for Computer Linguistics, 2014.

[63] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database systems - the complete book (2. ed.)*. Pearson Education, 2009.

[64] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9, 2006.

[65] Samira Ghodratnama, Amin Beheshti, Mehrdad Zakershahrak, and Fariborz Sobhanmanesh. Intelligent narrative summaries: From indicative to informative summarization. *Big Data Res.*, 26:100257, 2021.

[66] Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Stack: Sentence ordering with temporal commonsense knowledge. In *EMNLP*. Association for Computational Linguistics, 2021.

[67] Antonio Giuzio, Giansalvatore Mecca, Elisa Quintarelli, Manuel Roveri, Donatello Santoro, and Letizia Tanca. INDIANA: an interactive system for assisting database exploration. *Inf. Syst.*, 83:40–56, 2019.

[68] Dimitrios Gkesoulis, Panos Vassiliadis, and Petros Manousis. Cinecubes: Aiding data workers gain insights from OLAP queries. *Inf. Syst.*, 53:60–86, 2015.

[69] Dimos Gkitsakis, Spyridon Kaloudis, Eirini Mouselli, Verónika Peralta, Patrick Marcel, and Panos Vassiliadis. Cube interestingness: Novelty, relevance, peculiarity and surprise. *CoRR*, abs/2212.03294, 2022.

[70] Dimos Gkitsakis, Spyridon Kaloudis, Eirini Mouselli, Veronika Peralta, Patrick Marcel, and Panos Vassiliadis. Assessment methods for the interestingness of cube queries. In *DOLAP)*, volume 3369 of *CEUR Workshop Proceedings*, 2023.

[71] Dorota Glowacka, Evangelos E. Milios, Axel J. Soto, and Fernando Vieira Paulovich. Exploratory search and interactive data analytics. In *UI*, pages 9–11. ACM, 2017.

[72] Lukasz Golab and Divesh Srivastava. Exploring data using pa erns: A survey and open problems. In *DOLAP*, volume 2840 of *CEUR Workshop Proceedings*, pages 116–120. CEUR-WS.org, 2021.

[73] Matteo Golfarelli and Stefano Rizzi. A model-driven approach to automate data visualization in big data analytics. *Inf. Vis.*, 19(1), 2020.

[74] Ramanathan V. Guha, Vineet Gupta, Vivek Raghunathan, and Ramakrishnan Srikant. User modeling for a personal assistant. In Xueqi Cheng, Hang Li, Evgeniy Gabrilovich, and Jie Tang, editors, *WSDM*, pages 275–284. ACM, 2015.

[75] Hua Guo, Steven R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE Trans. Vis. Comput. Graph.*, 22(1):51–60, 2016.

[76] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Communications of the ACM*, 2012.

[77] David Holmes. *Communication Theory: Media, Technology, and Society.* SAGE Publications, 2005.

[78] Jiayi Hong. *Machine Learning Supported Interactive Visualization of Hybrid 3D and 2D Data for the Example of Plant Cell Lineage Specification. (Visualisation interactive, soutenue par l'apprentissage automatique, de données hybrides 3D et 2D ; l'exemple de la spécification du lignage cellulaire en biologie végétale).* PhD thesis, University of Paris-Saclay, France, 2023.

[79] Jessica Hullman and Nicholas Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2231–2240, 2011.

[80] Jessica Hullman, Steven Mark Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. A deeper understanding of sequence in narrative visualization. *IEEE TVCG*, 19(12):2406–2415, 2013.

[81] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. Improving comprehension of measurements using concrete re-expression strategies. In *CHI*, page 34. ACM, 2018.

[82] Jessica Hullman, Robert Kosara, and Heidi Lam. Finding a clear path: Structuring strategies for visualization sequences. *Comput. Graph. Forum*, 36(3):365–375, 2017.

[83] Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. Overview of data exploration techniques. In *SIGMOD*, pages 277–281. ACM, 2015.

[84] Manas Joglekar, Hector Garcia-Molina, and Aditya G. Parameswaran. Interactive data exploration with smart drill-down. *IEEE Trans. Knowl. Data Eng.*, 31(1):46–60, 2019.

[85] Jan Ole Johanssen, Anja Kleebaum, Bernd Bruegge, and Barbara Paech. How do practitioners capture and utilize user feedback during continuous software engineering? In *27th IEEE International Requirements Engineering Conference*, pages 153–164. IEEE, 2019.

[86] Petar Jovanovic, Sergi Nadal, Oscar Romero, Alberto Abelló, and Besim Bilalli. Quarry: A user-centered big data integration platform. *Inf. Syst. Frontiers*, 23(1):9–33, 2021.

[87] Jeongmi (Jamie) Kim and Daniel R. Fesenmaier. Measuring emotions in real time: Implications for tourism design. In *ENTER*. Springer, 2014.

[88] Jung-Hwan Kim, Minjeong Kim, Minjung Park, and Jungmin Yoo. Immersive interactive technologies and virtual shopping experiences: Differences in consumer perceptions between augmented reality (AR) and virtual reality (VR). *Telematics Informatics*, 77:101936, 2023.

[89] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus H. Gross, and Hanspeter Pfister. Visualizing nonlinear narratives with story curves. *IEEE Trans. Vis. Comput. Graph.*, 24(1):595–604, 2018.

[90] Maximilien Kintz. A semantic dashboard description language for a process-oriented dashboard design methodology. *Proceedings of MODIQUITOUS*, 2012.

[91] Robert Kosara. An argument structure for data stories. In *EuroVis*, pages 31–35. Eurographics Association, 2017.

[92] Robert Kosara and Jock Mackinlay. Storytelling: The next step for visualization. *IEEE Computer*, 46, 2013.

[93] Maha Ben Kraiem, Mohamed Alqarni, Jamel Feki, and Franck Ravat. OLAP operators for social network analysis. *Clust. Comput.*, 23(3):2347–2374, 2020.

[94] Sergey Kuznetsov, Alexey Tsyryulnikov, Vlad Kamensky, Ruslan Trachuk, Mikhail Mikhailov, Sergey Murskiy, Dmitrij V. Koznov, and George A. Chernishev. Unidata - A modern master data management platform. In *Proceedings of the Workshops of the EDBT/ICDT 2022 Joint Conference, Edinburgh, UK, March 29, 2022*, volume 3135 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[95] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. *Measuring User Engagement*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2014.

[96] Xingyu Lan, Xinyue Xu, and Nan Cao. Understanding narrative linearity for telling expressive time-oriented stories. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 604:1–604:13. ACM, 2021.

[97] Ana Lavalle, Alejandro Maté, Juan Trujillo, and Stefano Rizzi. Visualization requirements for business intelligence analytics: A goal-based, iterative framework. In *RE*, pages 109–119. IEEE, 2019.

[98] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5):84–90, 2015.

[99] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. Data-driven news generation for automated journalism. In *INLG*, pages 188–197. Association for Computational Linguistics, 2017.

[100] Wenyan Li, Alvin Grissom II, and Jordan L. Boyd-Graber. An attentive recurrent model for incremental prediction of sentence-final verbs. In *EMNLP*, volume EMNLP 2020 of *Findings of ACL*, pages 126–136. Association for Computational Linguistics, 2020.

[101] Lajanugen Logeswaran, Honglak Lee, and Dragomir R. Radev. Sentence ordering and coherence modeling using recurrent neural networks. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018.

[102] Pingchuan Ma, Rui Ding, Shi Han, and Dongmei Zhang. Metainsight: Automatic discovery of structured knowledge for exploratory data analysis. In *SIGMOD*, pages 1262–1274. ACM, 2021.

[103] Ruixian Ma, Honghui Mei, Huihua Guan, Wei Huang, Fan Zhang, Chengye Xin, Wenzhuo Dai, Xiao Wen, and Wei Chen. LADV: deep learning assisted authoring of dashboard visualizations from images and sketches. *IEEE Trans. Vis. Comput. Graph.*, 27(9):3717–3732, 2021.

[104] Paula Maddigan and Teo Susnjak. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and GPT-3 large language models. *IEEE Access*, 11:45181–45193, 2023.

[105] Shadan Malik. *Enterprise Dashboards: Design and Best Practices for IT*. John Wiley Sons, Inc., USA, 2005.

[106] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3):259–289, 1997.

[107] Patrick Marcel, Verónika Peralta, and Panos Vassiliadis. A framework for learning cell interestingness from cube explorations. In *ADBIS*, volume 11695 of *Lecture Notes in Computer Science*, pages 425–440. Springer, 2019.

[108] Nicholas Matthews. *Measurement, Levels of.* 2017.

[109] Mirjana Mazuran, Elisa Quintarelli, and Letizia Tanca. IQ4EC: intensional answers as a support to exploratory computing. In *DSAA*, pages 1–10. IEEE, 2015.

[110] Raymond Ondzigue Mbenga and et al. A data narrative about tuberculosis pandemic in gabon. In *Proceedings of the Workshops of the EDBT/ICDT*, volume 3135, 2022.

[111] David McCandless. *Information is Beautiful.* HarperCollins, 2012.

[112] Robert McKee. *Story: Substance, Structure, Style and the Principles of Screenwriting.* 1997.

[113] James D. McKeen, Heather A. Smith, and Satyendra Singh. Developments in practice XX - digital dashboards: Keep your eyes on the road. *Commun. Assoc. Inf. Syst.*, 16:52, 2005.

[114] Sean McKenna, Nathalie Henry Riche, Bongshin Lee, Jeremy Boy, and Miriah Meyer. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. *Comput. Graph. Forum*, 36(3):377–387, 2017.

[115] Florian Meier. Social network analysis as a tool for data analysis and visualization in information behaviour and interactive information retrieval research. In *CHIIR*, pages 477–480. ACM, 2020.

[116] Ronald A. Metoyer, Qiyu Zhi, Bart Janczuk, and Walter J. Scheirer. Coupling story to visualization: Using textual analysis as a bridge between data and interpretation. In *IUI*, pages 503–507. ACM, 2018.

[117] Tova Milo and Amit Somech. Deep reinforcement-learning framework for exploratory data analysis. In *SIGMOD*, pages 4:1–4:4. ACM, 2018.

[118] Tova Milo and Amit Somech. Next-step suggestions for modern interactive data analysis platforms. In *SIGKDD*, pages 576–585. ACM, 2018.

[119] Tova Milo and Amit Somech. Automating exploratory data analysis via machine learning: An overview. In *SIGMOD*, pages 2617–2622. ACM, 2020.

[120] Clément Moreau. *Fouille de séquences de mobilité sémantique : sur l'élaboration de mesures pour la comparaison, l'analyse et la découverte de comportements. (Mining Semantic Mobility Sequences: On the development of measures for comparison, analysis and discovery of discovery of behaviors).* PhD thesis, François Rabelais University, Tours, France, 2021.

[121] Gaia Mosconi, Dave Randall, Helena Karasti, Saja Aljuneidi, Tong Yu, Peter Tolmie, and Volkmar Pipek. Designing a data story: A storytelling approach to curation, sharing and data reuse in support of ethnographically-driven research. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW2):1–23, 2022.

[122] Nicholas Diakopoulos Sheelagh Carpendale Nathalie Henry Riche, Christophe Hurter. *Data-driven storytelling.* A K Peters/CRC Press, 2018.

[123] Samarth Navali, Jyothirmayi Kolachalam, and Vanraj Vala. Sentence generation using selective text prediction. *Computación y Sistemas*, 23(3), 2019.

[124] Cole Nussbaumer Knaflic. *Storytelling with Data: A Data Visualization Guide for Business Professionals.* John Wiley & Sons, 2015.

[125] Adegboyega Ojo and Bahareh Heravi. Patterns in award winning data storytelling: Story types, enabling tools and competences. *Digital Journalism*, 6, 2017.

[126] Carlos Ordonez, Zhibo Chen, and Javier García-García. Interactive exploration and visualization of OLAP cubes. In *DOLAP, pages = 83–88, publisher = ACM, year = 2011.*

[127] Faten El Outa, Matteo Francia, Patrick Marcel, Verónika Peralta, and Panos Vassiliadis. Supporting the generation of data narratives. In *ER Forum, Demo and Posters*, 2020.

[128] Faten El Outa, Matteo Francia, Patrick Marcel, Verónika Peralta, and Panos Vassiliadis. Towards a conceptual model for data narratives. In *ER*, pages 261–270, 2020.

[129] Faten El Outa, Patrick Marcel, and Verónika Peralta. Un modèle conceptuel de narration de données. In *Business Intelligence & Big Data, Actes de la conférence EDA 2021, en distanciel*, volume B-17 of *RNTI*, pages 49–54. Editions RNTI, 2021.

[130] Faten El Outa, Patrick Marcel, Verónika Peralta, Raphaël da Silva, Marie Chagnoux, and Panos Vassiliadis. Data narrative crafting via a comprehensive and well-founded process. In *Advances in Databases and Information Systems - 26th European Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings*, volume 13389 of *Lecture Notes in Computer Science*, pages 347–360. Springer, 2022.

[131] Themis Palpanas, Pawan Chowdhary, George A. Mihaila, and Florian Pinel. Integrated model-driven dashboard development. *Inf. Syst. Frontiers*, 9(2-3):195–208, 2007.

[132] Deok Gun Park, Mohamed Suhail, Minsheng Zheng, Cody Dunne, Eric D. Ragan, and Niklas Elmqvist. Storyfacets: A design study on storytelling with visualizations for collaborative data analysis. *Inf. Vis.*, 21(1):3–16, 2022.

[133] Dawn Cassandra Parker, Shahab Valaei Sharif, and Kaitlin Webber. Why did the ldquo;missing middlerdquo; miss the train? an actors-in-systems exploration of barriers to intensified family housing in waterloo region, canada. *Land*, 12(2), 2023.

[134] Peter Parycek and Markus Sachs. *Digital Storytelling: Capturing Lives, Creating Community*. Routledge, 2011.

[135] Koen Pauwels, Tim Ambler, Bruce Clark, Pat LaPointe, David Reibstein, Bernd Skiera, Berend Wierenga, and Thorsten Wiesel. Dashboards as a service : Why, what, how, and what research is needed? *Journal of Service Research*, 12:175–189, 2009.

[136] Eduardo H. M. Pena, Eduardo Cunha de Almeida, and Felix Naumann. Discovery of approximate (and exact) denial constraints. *Proc. VLDB Endow.*, 13(3):266–278, 2019.

[137] Protiva Rahman, Lilong Jiang, and Arnab Nandi. Evaluating interactive data systems. *VLDB J.*, 29(1):119–146, 2020.

[138] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *Proc. VLDB Endow.*, 10(11):1190–1201, 2017.

[139] Niklas Röber, Michael Böttinger, Bjorn Stevens, Amit Agrawal, and Francesca Samsel. Visualization of climate science simulation data. *IEEE Computer Graphics and Applications*, 41(1):42–48, 2021.

[140] Tuukka Ruotsalo, Jaakko Peltonen, Manuel J. A. Eugster, Dorota Glowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. Interactive intent modeling for exploratory search. *ACM Trans. Inf. Syst.*, 36(4):44:1–44:46, 2018.

[141] Sara Salem and Samir AbdelRahman. A multiple-domain ontology builder. In *COLING*, pages 967–975. Tsinghua University Press, 2010.

[142] Rushit Sanghrajka and R. Michael Young. Evaluating reader comprehension of plan-based stories containing failed actions. In *AIIDE*. AAAI Press, 2022.

[143] Sunita Sarawagi. User-adaptive exploration of multidimensional data. In *VLDB*, pages 307–316. Morgan Kaufmann, 2000.

[144] Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. Discovery-driven exploration of OLAP data cubes. In *EDBT*, volume 1377 of *Lecture Notes in Computer Science*, pages 168–182. Springer, 1998.

[145] Sunita Sarawagi and Gayatri Sathe. i^3: Intelligent, interactive investigaton of OLAP data cubes. In *SIGMOD*, page 589. ACM, 2000.

[146] Alper Sarikaya, Michael Correll, Lyn Bartram, Melanie Tory, and Danyel Fisher. What do we talk about when we talk about dashboards? *IEEE TVCG*, 25(1):682–692, 2019.

[147] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1139–1148, 2010.

[148] Jessica Zeitz Self, Radha Krishnan Vinayagam, J. T. Fry, and Chris North. Bridging the gap between user intention and model parameters for human-in-the-loop data analytics. In Carsten Binnig, Alan D. Fekete, and Arnab Nandi, editors, *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. ACM, 2016.

[149] D. Shi, F. Sun, X. Xu, Xingyu Lan, David Gotz, and Nan Cao. Autoclips: An automatic approach to video generation from data facts. *Comput. Graph. Forum*, 40(3):495–505, 2021.

[150] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Trans. Vis. Comput. Graph.*, 27(2):453–463, 2021.

[151] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Computer Society, 1996.

[152] Tarique Siddiqui, Surajit Chaudhuri, and Vivek R. Narasayya. COMPARE: accelerating groupwise comparison in relational databases for data analytics. *Proc. VLDB Endow.*, 14(11):2419–2431, 2021.

[153] Adalberto L. Simeone, Benjamin Weyers, Svetlana Bialkova, and Robert W. Lindeman, editors. *Everyday Virtual and Augmented Reality*. Human-Computer Interaction Series. Springer, 2023.

[154] Manish Singh, Michael J. Cafarella, and H. V. Jagadish. Dbexplorer: Exploratory search in databases. In *EDBT*, pages 89–100. OpenProceedings.org, 2016.

[155] Charles D. Stolper, Bongshin Lee, Nathalie Henry Riche, and John Stasko. Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories. Technical report, 2016.

[156] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. Erato: Cooperative data story editing via fact interpolation. *IEEE Trans. Vis. Comput. Graph.*, 29(1):983–993, 2023.

[157] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. Extracting top-k insights from multi-dimensional data. In *SIGMOD*, pages 1509–1524. ACM, 2017.

[158] Tan Tang, Sadia Rubab, Jiewen Lai, Weiwei Cui, Lingyun Yu, and Yingcai Wu. istoryline: Effective convergence to hand-drawn storylines. *IEEE Trans. Vis. Comput. Graph.*, 25(1):769–778, 2019.

[159] A. F. Thudt, C. Perin, W.C. Willett, and S. Carpendale. Subjectivity in personal storytelling with visualization. *Information Design Journal*, 23(1):48–64, 2017.

[160] Tzvetan Todorov. Les catégories du récit littéraire. *Communications*, 8:125–151, 1966.

[161] Panos Vassiliadis and Patrick Marcel. The road to highlights is paved with good intentions: Envisioning a paradigm shift in OLAP modeling. In *EDBT/ICDT*, volume 2062 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[162] Panos Vassiliadis, Patrick Marcel, Faten El Outa, Veronika Peralta, and Dimos Gkitsakis. A conceptual model for data storytelling highlights in business intelligence environments, 2024.

[163] Panos Vassiliadis, Patrick Marcel, and Stefano Rizzi. Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. *Inf. Syst.*, 85:68–91, 2019.

[164] Andrea Vázquez-Ingelmo, Francisco J. García-Peñalvo, and Roberto Therón. Information dashboards and tailoring capabilities - A systematic literature review. *IEEE Access*, 7:109673–109688, 2019.

[165] Rick Walker, Llyr ap Cenydd, Serban R. Pop, Helen C. Miles, Chris J. Hughes, William John Teahan, and Jonathan C. Roberts. Storyboarding for visual analytics. *Inf. Vis.*, 14(1):27–50, 2015.

[166] Qianwen Wang, Zhen Li, Siwei Fu, Weiwei Cui, and Huamin Qu. Narvis: Authoring narrative slideshows for introducing data visualization designs. *IEEE Trans. Vis. Comput. Graph.*, 25(1):779–788, 2019.

[167] Weiguo Wang, Sourav S. Bhowmick, Hui Li, Shafiq R. Joty, Siyuan Liu, and Peng Chen. Towards enhancing database education: Natural language generation meets query execution plans. In *SIGMOD , pages = 1933–1945, publisher = ACM, year = 2021.*

[168] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Trans. Vis. Comput. Graph.*, 26(1):895–905, 2020.

[169] Zezhong Wang, Harvey Dingwall, and Benjamin Bach. Teaching data visualization and storytelling with data comic workshops. In *CHI*. ACM, 2019.

[170] Wibke Weber, Martin Engebretsen, and Helen Kennedy. Data stories: Rethinking journalistic storytelling in the context of data journalism. *Studies in Communication Sciences*, 18:191–206, 2018.

[171] Steve Wexler, Jeffrey Shaffer, and Andy Cotgreave. *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. Wiley Publishing, 1st edition, 2017.

[172] Ryen W. White. *Interactions with Search Systems*. Cambridge University Press, 2016.

[173] Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. Goals, process, and challenges of exploratory data analysis: An interview study. *CoRR*, abs/1911.00568, 2019.

[174] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock D. Mackinlay, Bill Howe, and Jeffrey Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *CHI*, pages 2648–2659. ACM, 2017.

[175] Leni Yang, Xian Xu, Xingyu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. A design space for applying the freytag's pyramid structure to data stories. *IEEE Trans. Vis. Comput. Graph.*, 28(1):922–932, 2022.

[176] Zhihui Yang, Jiyang Gong, Chaoying Liu, Yinan Jing, Zhenying He, Kai Zhang, and X. Sean Wang. iexplore: Accelerating exploratory data analysis by predicting user intention. In *DASFAA018*, volume 10828 of *Lecture Notes in Computer Science*, pages 149–165. Springer, 2018.

[177] Shuainan Ye, Zhutian Chen, Xiangtong Chu, Yifan Wang, Siwei Fu, Lejun Shen, Kun Zhou, and Yingcai Wu. Shuttlespace: Exploring and analyzing movement trajectory in immersive visualization. *IEEE Trans. Vis. Comput. Graph.*, 27(2):860–869, 2021.

[178] Ogan M. Yigitbasioglu and Oana Velcu. A review of dashboards in performance management: Implications for design and research. *Int. J. Account. Inf. Syst.*, 13(1):41–59, 2012.

[179] Emanuel Zgraggen, Zheguang Zhao, Robert C. Zeleznik, and Tim Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *CHI*, page 479. ACM, 2018.

[180] Peiyan Zhang, Jiayan Guo, Chaozhuo Li, Yueqi Xie, Jaeboum Kim, Yan Zhang, Xing Xie, Haohan Wang, and Sunghun Kim. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *WSDM*, pages 168–176. ACM, 2023.

[181] Yangjinbo Zhang and Artur Lugmayr. Designing a user-centered interactive data-storytelling framework. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. Association for Computing Machinery, 2019.

[182] Yangjinbo Zhang, Mark Reynolds, Artur Lugmayr, Katarina Damjanov, and Ghulam Mubashar Hassan. A visual data storytelling framework. *Informatics*, 9(4), 2022.