Tutorial Paper

Data Narration for the People : Challenges and Opportunities

Patrick Marcel
University of Tours
Blois, France
Patrick.Marcel@univ-tours.fr

Veónika Peralta
University of Tours
Blois, France
Veronika.Peralta@univ-tours.fr

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes
Saint-Martin d'Hères, France
Sihem.Amer-Yahia@cnrs.fr

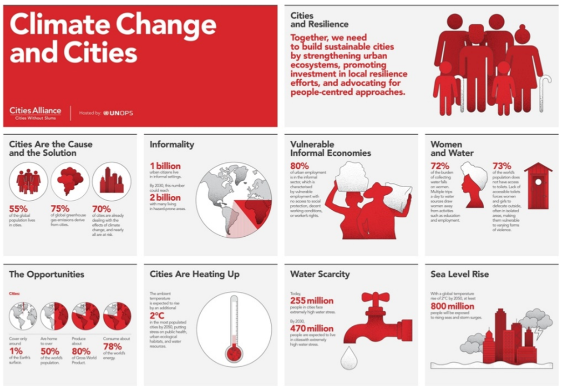

Climate Change and Cities
Cities Alliance
Cities Without Slums
Hosted by: UNOPS
Cities and Resilience
Together, we need to build sustainable cities by strengthening urban ecosystems, promoting investment in local resilience efforts, and advocating for people-centred approaches.
Cities Are the Cause and the Solution
55% of the global population lives in cities.
75% of global greenhouse gas emissions derive from cities.
70% of cities are already dealing with the effects of climate change, and nearly all are at risk.
Informality
1 billion urban citizens live in informal settings.
By 2030, this number could reach 2 billion with many living in hazard-prone areas.
Vulnerable Informal Economies
80% of urban employment is in the informal sector, which is characterised by vulnerable employment with no access to social protection, decent working conditions, or worker's rights.
Women and Water
72% of the burden of collecting water falls on women. Multiple trips a day to water sources draw women away from activities such as education and employment.
73% of the world's population does not have access to toilets. Lack of accessible toilets forces women and girls to defecate outside, often in isolated areas, making them vulnerable to varying forms of violence.
The Opportunities
Cities:
Cover only around 1% of the Earth's surface.
Are home to over 50% of the world's population.
Produce about 80% of Gross World Product.
Consume about 78% of the world's energy.
Cities Are Heating Up
The ambient temperature is expected to rise by an additional 2°C in the most populated cities by 2050, putting stress on public health, urban ecological habitats, and water resources.
Water Scarcity
Today, 255 million people in cities face extremely high water stress.
By 2030, 470 million people are expected to live in citieswith extremely high water stress.
Sea Level Rise
With a global temperature rise of 2°C by 2050, at least 800 million people will be exposed to rising seas and storm surges.

Figure 1: An example of data narrative

ABSTRACT

Data narration is the process of telling stories with insights extracted from data. It is an instance of data science [4] where the pipeline focuses on data collection and exploration, answering questions, structuring answers, and finally presenting them to stakeholders [16, 17]. This tutorial reviews the challenges and opportunities of the full and semi-automation of these steps. In doing so, it draws from the extensive literature in data narration, data exploration and data visualization. In particular, we point out key theoretical and practical contributions in each domain such as next-step recommendation and policy learning for data exploration, insight interestingness and evaluation frameworks, and the crafting of data stories for the people who will exploit them. We also identify topics that are still worth investigating, such as the inclusion of different stakeholders' profiles in designing data pipelines with the goal of providing data narration for all.

1 INTRODUCTION

Data narration [6, 23] refers to the notoriously tedious process of extracting insights from data and telling stories with the goal of "exposing the unanticipated" [27] and facilitating the understanding of insights. Data narration is practiced in many domains and by various domain experts, ranging from data journalists to public authorities. Figure 1 shows an example of a data narrative about climate change[1].

According to De Bie et al. [4], data narration poses the greatest challenges for automation, since background knowledge and human judgment are key to its success. In this tutorial, we review the state of the art in automating data narration and the challenges and perspectives that arise from that.

In particular, we consider the process of data narration as an instance of data science pipelines (see Figure 2) [17] that start with (i) *exploring data*, a cumbersome phase including the collection, preparation and analysis of data, whose aim is to extract insights, followed by (ii) *answering questions*, a phase where the narrator derives from insights the messages that they intend to communicate as answers to analytical questions or to an analysis goal, (iii) *structuring answers* for organizing messages into story episodes, and finally (iv) *presenting messages* via visual artifacts that can be easily communicated to an intended audience. We present related work on formalisms, languages and algorithms to automate these steps and evaluate their outcomes. Throughout the presentation, we focus on the stakeholders who make use of the narratives and ask the question of what is missing to facilitate their job.

2 OUTLINE

The tutorial is divided into three main parts, as follows:

- **Part 1: Data narration [20mn]**
  1. *Use cases and Applications*: presentation of various use cases ranging from data journalism to worldwide health situation monitoring, with an emphasis on what is automated.
  2. *Conceptual Model*: clarification of the concepts and terminology behind data narration, and introduction of the data narrative crafting process.
- **Part 2: Approaches [40mn]**
  1. *Evaluation*: review of approaches and benchmarks to evaluate a narrative as well as each phase of data narration.
  2. *Present*: review of visualization techniques.
  3. *Structure Answers*: presentation of classical data narrative structures, and review of manual approaches for structuring the story plot.
  4. *Answer Questions*: review of approaches for modeling the user's intentions in devising a story.
  5. *Explore*: review of manual, partially guided and fully guided exploration approaches.
- **Part 3: Perspectives [20mn]** Discussion of perspectives in all the four phases (Explore, Answer Questions, Structure Answers, Present), with a particular attention to the inclusion of different stakeholders' profiles in designing data pipelines.

[1] Cities Alliance: "Climate Change and Cities – Infographic". https://www.citiesalliance.org/newsroom/news/results/climate-change-and-cities-infographic

Series ISSN: 2367-2005

10.48786/edbt.2023.82

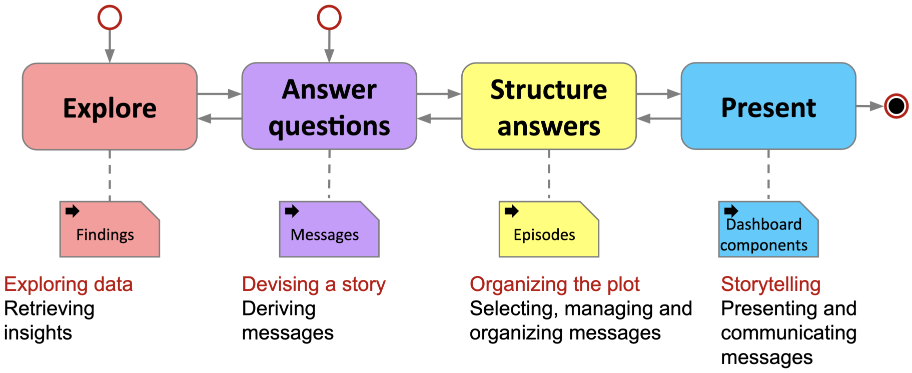


Figure 2: Data Narration as a data science pipeline

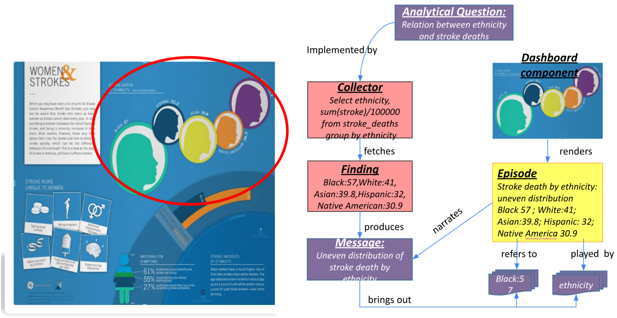

WOMEN & STROKES
Analytical Question:
Relation between ethnicity and stroke deaths
Implemented by
Collector
Select ethnicity, sum(stroke)/100000 from stroke_deaths group by ethnicity
fetches
Finding
Black:57,White:41, Asian:39.8,Hispanic:32, Native American:30.9
produces
Message:
Uneven distribution of stroke death by ethnicity
narrates
Dashboard component
renders
Episode
Stroke death by ethnicity: uneven distribution Black 57 ; White:41; Asian:39.8; Hispanic: 32; Native America 30.9
refers to
played by
Black:57
ethnicity
brings out

Figure 3: Concepts of a data narrative

Part 1 introduces data narration, its concepts [16], and its use in various applications. It clarifies the terminology around data narration, borrowing from narrative theory [8] where a narrative consists of a story (the content) and a discourse (the expression). Running examples are used to illustrate the concepts and terminology (see Figure 3). This part then sheds light on what is automated in data narrative crafting [17]. This is particularly important as data narration is tedious and it is not always easy to distinguish between labor intensive tasks and easily automated ones.

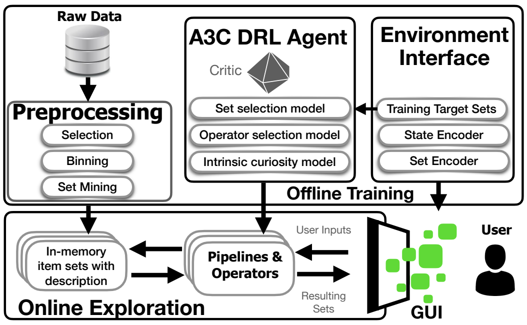


Figure 4: Overall Architecture of DORA The Explorer

Part 2 addresses automation by breaking down the data narration pipeline (the four phases of Figure 2) and distinguishing existing semi- to fully automated techniques.

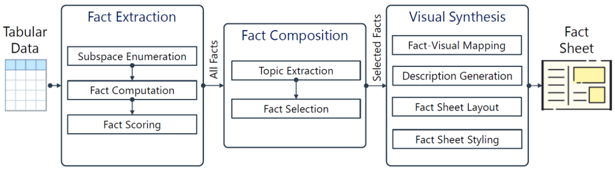


Figure 5: Overall Architecture of Datashot

We start by reviewing existing evaluation frameworks and present a comprehensive description of evaluation metrics that we organize into Human, System and Data metrics, following [18]. Figure 7 shows a snapshot of different metrics. It extends the framework proposed in [21] that categorizes evaluation metrics used for a data exploration system into System and Human metrics. The former mainly focus on capturing response time (e.g., time delays in query scheduling and processing), while the latter focus on quantifying human behavior (e.g. exploration duration) and satisfaction by deploying user surveys. The Data metrics relate to the quality of returned answers and narratives. We see how existing frameworks focus on different metrics. We review existing works in evaluating data visualization and data exploration. We show that benchmarks proposed by the data management community [3, 9] remain mostly system-centric and fail to capture the complexity of data narration. The goal of these studies is to evaluate the query processing engines, but they do not account for user perception in terms of, e.g., engagement or understandability.

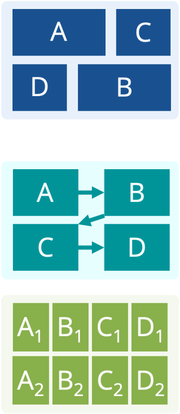


Figure 6: Structures of fact sheets: random, sequential and multiple series (from [29])

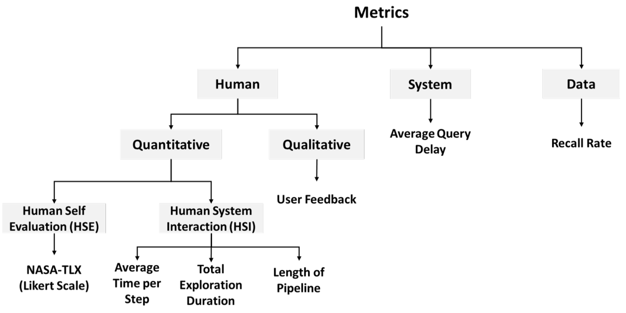


Figure 7: A snapshot of quantitative and qualitative metrics.

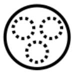
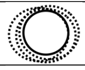
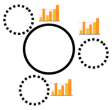
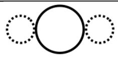
Operator	RCC8 Formalism [22]	Output description
by-facet(D, A)	NTPPi 	returns as many subsets of D as there are combinations of values of attributes in A
by-superset(D, k)	NTPP 	returns the k smallest supersets of input set D (k is application-dependent)
by-distribution(D)	DC 	returns all sets that are distinct from the input set D and whose attribute value distribution is similar to D
by-neighbors(D, a)	EC 	returns 2 sets that are distinct from the input set D and that have the previous (smaller) and next (larger) values for attribute a

Table 1: Exploration operators. The second column illustrates the equivalent definition of each operator in the *Region Connection Calculus 8* (RCC8) formalism [13, 22]. The input data is represented with a bold line and the output results are represented with dashed lines.

Present reviews work on communicating insights in a creative way [25]. We describe existing systems that automate visualization, including the visualization modules of systems implementing all the data narration pipeline, introduced before [24, 26, 29]. Various applications are used as illustrations, ranging from exploring galaxies [19] to summarizing music tracks [30].

Structure answers starts with a review of classical data narrative structures, from the basic organization of data sheets (see e.g., Figure 6) to interactivity-based design [23]. We describe some existing systems that automate data structuring, such as Calliope [24] and Erato [26], and include a focus on transitions in the story ([11, 24]).

Answer questions includes a focus on data narrative intents [2] and how users can express their intentions using exploration primitives [20], insight specific primitives [7] or high level intentional languages [28]. For instance, trained policies in DORA The Explorer make use of operators summarized in Table 1.

Explore includes a focus on insights' interestingness [14]. We present a description of some existing systems that automate data exploration such as DORA The Explorer [20], ATENA [10] and Datashot [29]. For instance, we describe the architecture of DORA The Explorer (see Figure 4) [20], which consists of an offline phase to train exploration models, using a Tensorflow-based implementation of A3C[2], and an online phase to deploy trained exploration models. We also review the data exploration modules of systems implementing all the data narration pipeline. For instance, we describe the first two modules of Datashot (see Figure 5) [29] that generate insights.

Part 3 describe open research perspectives in each step of the data narration pipeline, as well as challenges for the building of an end-to-end data narration system. A particular focus is made on describing human-in-the-loop data narration with individual and collaborative aspects [1], the possibility to use the VALIDE framework [18] for evaluating narratives, as well as the ability to express questions as hypotheses and produce statistically sound narratives [5].

3 GOALS AND OBJECTIVES

The objective of the tutorial is to draw a landscape of recent advances in data narration and their relationship with data science pipelines. It helps the audience distinguish between labor intensive and easily automated tasks of data exploration, answering questions, structuring results, and presenting and visualizing insights. This landscape allows us to raise research challenges and opportunities in making data narration accessible to all.

4 INTENDED AUDIENCE

The tutorial targets researchers and postgraduate students interested in Data Science, and particularly Data Narration, Data Visualization and Interactive Data Exploration. Attendees should have a background in database and information systems.

5 BIOGRAPHY

Patrick Marcel <*patrick.marcel@univ-tours.fr*> is an Associate Professor at the University of Tours, France. His current research focuses on OLAP and data warehousing, recommender systems, exploratory data analysis and data narration. Patrick served as program committee member in top tier international conferences, including ER, VLDB, EDBT. He is a member of the steering committee of DOLAP and a member of the regular editorial board of DKE.

Verónika Peralta <*veronika.peralta@univ-tours.fr* is an Associate Professor at the University of Tours (France) where she is head of the Computer Science department. She received her Ph.D. in 2006 from the University of Versailles (France) and the University of the Republic (Uruguay). Her current research interests include data and information quality, exploratory data analysis, business intelligence and data narration. She served as program committee member and guest editor in many international conferences and journals.

Sihem Amer-Yahia <*sihem.amer-yahia@cnrs.fr*> is a Silver Medal CNRS Research Director and Deputy Director of the Lab of Informatics of Grenoble. She works on exploratory data analysis and fairness in job marketplaces. Before joining CNRS, she was Principal Scientist at QCRI, Senior Scientist at Yahoo! Research and Member of Technical Staff at at&t Labs. Sihem is PC chair of ACM SIGMOD 2023. She leads the Diversity, Equity and Inclusion initiative for the DB community.

6 EARLIER VERSION OF THE TUTORIAL

This present tutorial can be seen as a follow-up to earlier tutorial on data exploration [12, 15], in that it considers data exploration in the general context of data narration.

[2] https://github.com/marload/DeepRL-TensorFlow2/

Part 1, and a few elements of Part 2 of this tutorial were presented at e-EGC 2022[3], a winter school joined with EGC2022, the French conference on data management and discovery[4], and at eBISS 2022, an international summer school in big data management and analytics[5]. The version presented at eBISS 2022 can be accessed online[6].

The focus of this new tutorial is on bridging the gap between data narration and large-scale data exploration to make narration accessible to all stakeholders. The material of the previous tutorial is reused in Part 1 and at some point in Part 2 of the new tutorial. Most of Part 2 and Part 3 consist of new material.

REFERENCES

[1] Sihem Amer-Yahia, Shady Elbassuoni, Behrooz Omidvar-Tehrani, Ria Mae Borromeo, and Mehrdad Farokhnejad. Grouptravel: Customizing travel packages for groups. In *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, pages 133–144, 2019.

[2] 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*, chapter 5, pages 107–133. Taylor Francis, 2018.

[3] Leilani Battle, Philipp Eichmann, Marco Angelini, Tiziana Catarci, Giuseppe Santucci, Yukun Zheng, Carsten Binnig, Jean-Daniel Fekete, and Dominik Moritz. Database benchmarking for supporting real-time interactive querying of large data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1571–1587, 2020.

[4] Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, and Christopher K. I. Williams. Automating data science. *Commun. ACM*, 65(3):76–87, 2022.

[5] Nassim Bouarour, Idir Benouaret, and Sihem Amer-Yahia. Significance and coverage in group testing on the social web. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3052–3060. ACM, 2022.

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

[7] Alexandre Chanson, Nicolas Labroche, Patrick Marcel, Vincent T'Kindt, and Stefano Rizzi. Automatic generation of comparison notebooks for interactive data exploration. In *EDBT*, 2022.

[8] S.B. Chatman. *Story and Discourse: Narrative Structure in Fiction and Film.* 1980.

[9] Philipp Eichmann, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. Idebench: A benchmark for interactive data exploration. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1555–1569, 2020.

[10] Ori Bar El, Tova Milo, and Amit Somech. Automatically generating data exploration sessions using deep reinforcement learning. In *SIGMOD*, 2020.

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

[12] Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. Overview of data exploration techniques. In *SIGMOD*, 2015.

[13] Sanjiang Li and Mingsheng Ying. Region connection calculus: Its models and composition table. *Artificial Intelligence*, 145(1-2):121–146, 2003.

[14] Patrick Marcel, Verónika Peralta, and Panos Vassiliadis. A framework for learning cell interestingness from cube explorations. In *ADBIS*, 2019.

[15] Tova Milo and Amit Somech. Automating exploratory data analysis via machine learning: An overview. In *SIGMOD*, 2020.

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

[17] 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*, pages 347–360, 2022.

[18] Yogendra Patil, Sihem Amer-Yahia, and Srividya Subramanian. Designing the evaluation of operator-enabled interactive data exploration in VALIDE. In *HILDA@SIGMOD*, pages 4:1–4:7. ACM, 2022.

[19] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, and Srividya Subramanian. Balancing familiarity and curiosity in data exploration with deep reinforcement learning. In *aiDM '21: Fourth Workshop in Exploiting AI Techniques for Data Management, Virtual Event, China, 25 June, 2021*, pages 16–23, 2021.

[20] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, and Srividya Subramanian. DORA THE EXPLORER: exploring very large data with interactive deep reinforcement learning. In *CIKM*, 2021.

[21] Lilong Jiang Protiva Rahman and Arnab Nandi. Evaluating interactive data systems - survey and case studies. *The VLDB Journal.*, 29:119–146, 2020.

[22] David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, KR'92, page 165–176, 1992.

[23] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *TVCG*, 16(6):1139–1148, 2010.

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

[25] 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.

[26] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. Erato: Cooperative data story editing via fact interpolation. *CoRR*, abs/2209.02529, 2022.

[27] John W. Tukey. *Exploratory data analysis.* Addison-Wesley series in behavioral science : quantitative methods. 1977.

[28] 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.

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

[30] Brit Youngmann, Sihem Amer-Yahia, and Aurélien Personnaz. Guided exploration of data summaries. *Proc. VLDB Endow.*, 15(9):1798–1807, 2022.

[3] https://egc2022.univ-tours.fr/ecole/,https://egc2022.univ-tours.fr/programme-e-egc/

[4] https://egc2022.univ-tours.fr/

[5] https://cs.ulb.ac.be/conferences/ebiss2022/index.html

[6] https://cs.ulb.ac.be/conferences/ebiss2022/slides/marcel_peralta_1.pdf, https://cs.ulb.ac.be/conferences/ebiss2022/slides/marcel_peralta_2.pdf