Datastories

BIG DATA

FUNDAMENTALS AND NARRATIVES

Enhancing Skills for European University Students and Academics

V1.0_Feb2024

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.
Project ID: 2021-1-SE01-KA220-HED-00032034

Project partners:

KTH Royal Institute of Technology, Sweden

Vytautas Magnus University, Lithuania

Thomas More University of Applied Sciences, Belgium

i-strategies, Italy

MEET Digital Culture Center, Italy

Project webpage: https://www.bigdatastoryteller.eu/

Authors: Ceci, A. (i-Strategies), De Ridder, K. (TM), Dinuzzo, R. (MEET), Hedman, A. (KTH), Krilavičius, T. (VDU), Li, H. (KTH), Scholtz, M. (TM), Stornaiuolo, L. (Toretei), Vagnarelli, G. (i-strategies), Vita, T. (MEET), Zdanavičiūtė, M. (VDU),

ISBN: 978-91-8040-955-1

Cover page images source: Created by DALL-E 2, an AI system designed by OpenAI (2024)

"Every bit of the data universe tells a story."

- *ChatGPT 4, OpenAI, 2024.*

Acknowledgements

The project partners would like to thank the following institutions and organisations for their contributions to this project:

1. The European Commission's Erasmus+ funding programme, for its financial contribution that made this project possible.

2. All the organisations and experts who contributed their time and expertise to achieve the objectives and goals of the EU Erasmus+ Datastories project:

BELGIUM

Kaat De Ridder from "Thomas More University of Applied Sciences"
Marco Scholtz from "Thomas More University of Applied Sciences"

ITALY

Alessio Ceci from "i-strategies"
Gianluca Vagnarelli from "i-strategies"
Luca Stornaiuolo from "MEET"
Rosanna Dinuzzo from "MEET"
Thelma Vita from "MEET"

LITHUANIA

Aušra Mackutė-Varoneckienė from "Vytautas Magnus University"
Monika Zdanavičiūtė from "Vytautas Magnus University"
Tomas Krilavičius from "Vytautas Magnus University"

SWEDEN (project coordinator)

Anders Hedman from "KTH Royal Institute of Technology"
Haibo Li from "KTH Royal Institute of Technology"

Table of contents

What is this guidebook about, and how should I use it? xi

1. Purpose of this guidebook xi

2. Target Audience xi

3. How to Use This Guidebook xii

Big Data datasets xiii

Big Data webinars xiv

i. Webinars for Chapter 1: The Importance of Big Data Storytelling xiv

a) An introduction to Big Data storytelling xiv

b) Big Data storytelling for social change xiv

c) EU Strategy on Big Data xiv

ii. Webinars for Chapter 2: The Unreasonable Effectiveness of Big Data xv

a) The unreasonable effectiveness of Big Data xv

iii. Webinars for Chapter 3: AI Storytelling – A Human-Centered Perspective from Moral Philosophy xv

a) Big Data storytelling: Ethical & human-centred considerations xv

iv. Webinars for Chapter 4: The Visualisation of Big Data xv

a) Visualisation methods for Big Data Storytelling xv

b) Multidimensional data visualisation for Big Data Storytelling: Direct methods xv

c) Multidimensional data visualisation for Big Data Storytelling: Projections xvi

D) Tableau & Data visualisation: A simple Big Data tool xvi

v. Webinars for Chapter 5: Story-based Data Analytics xvi

a) Story-driven data analytics xvi

vi. Webinars for Chapter 6: The case of Big Data in intangible cultural heritage tourism xvii

a) Data visualisation for intangible cultural heritage: An introduction xvii

b) The role of Big Data in tourism and storytelling xvii

c) Big Data analyses for ICH Tourism storytelling xvii

vii. Webinars for Chapter 7: The Future of Big Data Storytelling Techniques xviii

a) The future of Data Storytelling 1: Introduction to AI-powered Data Storytelling Techniques xviii

b) The future of Data Storytelling 2: The technical aspects of AI-powered Data Storytelling xviii

c) The future of Data Storytelling 3: Applications and Case Studies xviii

Chapter 1: The Importance of Big Data Storytelling 1

1.1 What is Big Data 1

1.1.1 The five "V" of Big Data 3

1.2 Big Data Storytelling 4

1.2.1 We live in the narrative era 5

1.2.2 Big Data Storytelling 7

1.2.3 Three examples of data storytelling 8

1.3 Big Data for Social Change 14

1.3.1 Big Data for Social Change Meaning 15

1.3.2 Practical Examples of Big Data for Social Change 16

1.4 The European Strategy for Data 19

1.5 Conclusion 19

Chapter 2: The Unreasonable Effectiveness of Big Data 21

2.1 Introduction 21

2.2 The Effectiveness of Big Data 22

2.2.1 The Unexpected Association 22

2.3 Effectiveness in Vision Tasks 23

2.4 A Mathematical Explanation of The Unreasonable Effectiveness of Big Data 24

2.4.1 Birthday Problem 25

2.4.2 Same Birthday as Yours 27

2.5 Combinational Exploration and Spurious Relation 30

2.6 Concluding Remarks 30

Chapter 3: AI Storytelling – A Human-Centered Perspective from Moral Philosophy 32

3.1 Intentionality in humans and machines 32

3.1.1 The Chinese room thought experiment 33

3.1.2 Syntax and Semantics 34

3.2 Human Storytelling 34

3.3 Machine Storytelling 35

3.3.1 Fixing hallucinations 35

3.3.2 Common sense 35

3.3.3 Wisdom 35

3.4 Hybrid Storytelling 36

3.4.1 Machine-dominant contribution 36

3.4.2 Equal contribution 36

3.4.3 Human-dominant contribution 36

3.5 Authentic Storytelling 37

3.5.1 Ownership 37

3.5.2 Human storytelling and authenticity 37

3.5.3 Machine storytelling and authenticity 37

3.5.4 Hybrid storytelling and authenticity 37

3.6 Conclusion 38

3.7 Study Questions 38

Chapter 4: The Visualisation of Big Data **39**

4.1. Basic visualisation methods 40

4.1.1. Basic (Line) charts 40

4.1.2. Area Charts 42

4.1.3. Bar charts 43

4.1.4. Histograms 48

4.1.5. Pie charts 49

4.1.6. Box plots 51

4.2 Multidimensional data storytelling: direct methods 52

4.2.1 Scatter plot 54

4.2.2 Heat maps 57

4.2.3 Graphs 58

4.2.4 Mosaic plot 59
4.2.5 Sankey 60
4.2.6 Parallel coordinates 61
4.3 Multidimensional data storytelling: projections 62
4.3.1. Principal Components Analysis 63
4.3.2 Multidimensional scaling 65
4.4 Data visualisation tools/software 67
Chapter 5: Story-based Data Analytics 68
5.1 Introduction 68
5.2 A Creative Data-driven Approach 69
5.3 Leading with Decision-driven Data Analytics 71
5.4 Story-driven Approach to Information Exploration 73
5.4.1 The Wicked Problem 73
5.4.2 Storytelling 74
5.4.3 Data Analytics Driven by Story 75
5.5 Concluding Remarks 76
Chapter 6: The Case of Big Data in intangible cultural heritage tourism (TM) 79
6.1 Intangible cultural heritage and tourism 79
6.1.1 Why is the tourism industry important? 80
6.1.2 Intangible Cultural Heritage (ICH) Tourism 80
6.1.3 UNESCO’s 2003 Convention for ICH 82
6.2 The use of data in tourism 82
6.2.1 The significance of tourism data 83
6.2.2 Limitations of Traditional Tourism Datasets 84
6.3 Big Data in Tourism 85
6.3.1 The use of Big Data in Tourism 85
6.3.2 The kinds of Big Data collected in Tourism 87
6.3.3 Businesses/organisations that make use of Big Data 89
6.4 Big Data in CHT 93

6.4.1 Big Data storytelling in ICH – A practical example from UNESCO 93
6.4.2 Big Data Storytelling in Tourism – practical example from UNTWO 95
6.4.3 Big Data Storytelling in ICH CHT – self-created example 97
6.5 Conclusion 100
Chapter 7: The Future of Big Data Storytelling Techniques 101
7.1 Emerging Technologies in Data Visualisation 101
7.1.1 Augmented Reality (AR) and Virtual Reality (VR) in Data Visualisation 101
7.1.2 Artificial Intelligence (AI) and Machine Learning (ML) in Data Visualisation 101
7.1.3 Real-time Data Visualisation 102
7.1.4 Interactive Data Visualisation 102
7.1.5 Collaborative Data Visualisation 102
7.2 Artificial Intelligence (AI) and Machine Learning 103
7.2.1 Automated Insight Generation 103
7.2.2 Predictive Analytics and Forecasting 103
7.2.3 Enhancing Data Visualisation with ML 103
7.2.4 Interactive and Dynamic Visualisations 104
7.2.5 Ethical Considerations and Bias Mitigation 104
7.2.6 Future Trends 104
7.3 Ethical Considerations in Data Storytelling 105
7.3.1 Data Privacy Concerns 105
7.3.2 Addressing Bias and Ensuring Fairness 106
7.3.3 Future Directions 106
7.4 Interdisciplinary approaches to Data Science 107
7.4.1 Journalism and Data Science 107
7.4.2 Visual Arts and Data Visualisation 107
7.4.3 Psychology and User Experience (UX) Design 108
7.4.4 Computer Science and Interactive Technologies 108
7.4.5 Ethnography and Cultural Studies 108
7.4.6 Future Directions 108

7.5 Personalisation and audience engagement 109
7.5.1 Strategies for Personalised Storytelling 109
7.5.2 Technologies Enabling Personalised Storytelling 110
7.5.3 Examples of Personalised Data Storytelling 110
7.5.4 Future Directions 111
7.6 Future Challenges 111
7.6.1 Future Opportunities 112
7.6.2 Navigating the Future 113
Reference lists **114**
Chapter 1 references 114
Chapter 2 references 116
Chapter 3 references 116
Chapter 4 references 116
Chapter 5 references 117
Chapter 6 references 117
Chapter 7 references 119

What is this guidebook about, and how should I use it?

1. Purpose of this guidebook

This guidebook provides the fundamental tools and know-how to use Big Data and data visualisation to tell compelling stories. Our goal? To help you dive into those tricky social issues and spark creative problem-solving. Plus, we've covered you with real-world examples from the world of Intangible Cultural Heritage (ICH) tourism – so you can see how data storytelling works in action.

2. Target Audience

This guidebook is designed to cater to a diverse audience interested in enhancing their ability to convey meaningful narratives using data. However, its design is mainly for the following two primary groups:

- **University Students:** This guidebook is valuable for students pursuing degrees in data science, communication, or related fields. It equips them with the essential skills to navigate the realm of Big Data and transform it into compelling narratives.
- **Educators and Lecturers:** For educators, lecturers, and course instructors, this guidebook offers a structured approach to integrating Big Data visualisation into their curriculum. It provides a comprehensive foundation for effectively teaching data storytelling techniques.

Although this guidebook is created with students and educators in mind, the following groups can also benefit from this document:

- **Researchers and Practitioners:** Researchers and professionals engaged in data analysis and storytelling will find this guidebook beneficial in expanding their expertise. It offers insights into best practices, ethical considerations, and real-world examples.
- **Professionals across Industries:** Individuals working in diverse industries can derive practical guidance from this guidebook. It offers valuable insights into using data visualisation to address societal challenges, regardless of the sector.
- **Storytellers and Communicators:** This guidebook will be invaluable to storytellers, communicators, and content creators who aim to enhance their narrative capabilities through data-driven insights.

As you explore this guidebook, consider your unique perspective and how it can assist you in harnessing the power of Big Data to craft compelling narratives and drive positive change.

3. How to Use This Guidebook

This guidebook is designed to cater to a wide range of readers, from beginners to those with some background in Big Data and storytelling. Whether seeking a comprehensive understanding or specific insights, you can navigate the guidebook efficiently based on your needs. The following points will act as a guide on how this document will best serve your purposes.

i. For Novice Readers or Beginners

If you are new to Big Data and storytelling, we recommend starting from the beginning and following the chapters sequentially. This approach will provide you with a structured foundation, building your knowledge step by step.

ii. For Readers with Some Background

Suppose you already have a basic understanding of Big Data and storytelling. In that case, you may skip directly to the chapters that align with your interests or areas where you wish to deepen your knowledge.

iii. Chapter Directory: Finding What You Need

Use the following directory to locate the chapters that address your specific interests quickly:

i. I want to learn about the **basics of Big Data**, its importance, application, and transformative impact across various sectors: **Chapter 1**

ii. I'm interested in the **mathematical exploration** of Big Data to grasp data analytic systems better, aiming to **revolutionise problem-solving and decision-making**: **Chapter 2**

iii. I want to explore **human-centred storytelling** (human, machine, and hybrid storytelling), touching on treating **AI as storytellers**, and focusing on generative AIs despite their limited storytelling capabilities: **Chapter 3**

iv. I want to **learn about the different methods** in which I can **illustrate/visualise Big Data** to which I can apply s**torytelling**: **Chapter 4**

v. I want to learn more about **story-driven data analytics** by exploring a **collaborative and iterative approach** for **complex problem solving**: **Chapter 5**

vi. I want to **examine a case study** (intangible cultural heritage tourism) regarding the importance and use of Big Data, together with storytelling **examples**: **Chapter 6**

vii. I want to see what **the future of Big Data storytelling** looks like: **Chapter 7**

Big Data datasets

During the project, various Big Data datasets were collected and uploaded or linked to the project's online repository with the purpose of being used as examples/inspiration while learning about Big Data. These datasets are openly available, thus feel free to use them. You can access these datasets on the project website by following the link below.

ACCESS THE DATASETS

Big Data webinars

There are accompanying webinars for all the content of this book. In this section, you will find additional information regarding the recorded webinars to help explain the contents of this guidebook. You can go through the information below to see what information you are interested in. These are coupled with the various chapters in this book.

i. Webinars for Chapter 1: The Importance of Big Data Storytelling

a) An introduction to Big Data storytelling

This first webinar provides a glimpse into how Big Data has revolutionised the way we live while also clarifying what Big Data encompasses, the benefits it provides, and providing examples that help us understand it better.

START WEBINAR	**28:09** *Runtime*	**Mr Gianluca Vagnarelli** *Profound Innovation and Social Responsibility* *i-strategies, Offida, Italy* *Author*

b) Big Data storytelling for social change

We delve into the transformative power of Big Data storytelling in tackling societal issues, showcasing its utility in fostering equity, justice, and proactive social reforms through activism, philanthropy, and social justice efforts.

START WEBINAR	**30:02** *Runtime*	**Mr Gianluca Vagnarelli** *Profound Innovation and Social Responsibility* *i-strategies, Offida, Italy* *Author*

c) EU Strategy on Big Data

In this webinar we further explore the transformative power of Big Data, while also zooming in on the EU approach and initiatives regarding Big Data.

START WEBINAR	**22:05** *Runtime*	**Mr Alessio Ceci** *Profound Innovation and Social Responsibility* *i-strategies, Offida, Italy* *Author*

ii. Webinars for Chapter 2: The Unreasonable Effectiveness of Big Data

a) The unreasonable effectiveness of Big Data

In this webinar, employing the birthday problem as a conceptual tool, we explore the mathematical dimensions of numerous challenging yet connected questions in big data: why do unexpected associations arise? what does the term 'unreasonable effectiveness' imply? Why is this effectiveness considered unreasonable? and why do spurious correlations emerge?

START WEBINAR	**16:48** *Runtime*	**Prof. Haibo Li** *KTH Royal Institute of Technology, Sweden* *Author*

iii. Webinars for Chapter 3: AI Storytelling – A Human-Centered Perspective from Moral Philosophy

a) Big Data storytelling: Ethical & human-centred considerations

This webinar considers different forms of AI-supported storytelling and the question of authentic storytelling.

START WEBINAR	**13:30** *Runtime*	**Dr Anders Hedman** *KTH Royal Institute of Technology, Sweden* *Author*

iv. Webinars for Chapter 4: The Visualisation of Big Data

a) Visualisation methods for Big Data Storytelling

This webinar presents basic visualisation methods for big data storytelling. It describes the following methods and their different variations: line chart, area graph, bar chart, Cleveland dot plot, histogram, pie chart, box plot, and error bar.

START WEBINAR	**19:23** *Runtime*	**Prof Tomas Krilavičius** **Ms Monika Zdanavičiūtė** *Faculty of Informatics at Vytautas Magnus University, Kaunas, Lithuania* *Authors*

b) Multidimensional data visualisation for Big Data Storytelling: Direct methods

This webinar presents ways of using multidimensional data and direct visualisation methods for big data storytelling. The direct methods mean that all attributes of the multidimensional object are presented visually. Examples of such methods include scatter plots, trellis, heat maps, graphs, trees, mosaic plots, Sankey, and parallel coordinates.

[START WEBINAR] **25:53** *Runtime*

Prof Tomas Krilavičius
Ms Monika Zdanavičiūtė
Faculty of Informatics at Vytautas Magnus University, Kaunas, Lithuania
Authors

c) Multidimensional data visualisation for Big Data Storytelling: Projections

This webinar presents ways of using multidimensional data and direct visualisation methods for big data storytelling. The direct method means that all attributes of the multidimensional object are presented visually. Such methods include scatter plots, trellis, heat maps, graphs, trees, mosaic plots, sankey, and parallel coordinates.

[START WEBINAR] **20:07** *Runtime*

Prof. Tomas Krilavičius
Ms Monika Zdanavičiūtė
Faculty of Informatics at Vytautas Magnus University, Kaunas, Lithuania
Authors

D) Tableau & Data visualisation: A simple Big Data tool

Through this webinar, we explore the software known as Tableau, as well as other forms of Big Data visualisation software, which makes Big Data easier to display and interpret.

[START WEBINAR] **18:42** *Runtime*

Ms Ilenia Chiodi
i-strategies, Offida, Italy
Author

v. Webinars for Chapter 5: Story-based Data Analytics

a) Story-driven data analytics

In this webinar, we present a new approach to data analytics called story-driven data analytics. This approach is notable for its focus on storytelling and collaboration between technical and business teams. Unlike other approaches, such as data-driven and decision-oriented data analytics, it is ideal for complex problem-solving situations requiring diverse perspectives and expertise.

[START WEBINAR] **16:33** *Runtime*

Prof. Haibo Li
KTH Royal Institute of Technology, Sweden
Author

vi. Webinars for Chapter 6: The case of Big Data in intangible cultural heritage tourism

a) Data visualisation for intangible cultural heritage: An introduction

In this webinar, we delve into the meaning of heritage, specifically focusing on Intangible Cultural Heritage (ICH). We will examine the UNESCO Convention for Safeguarding the Intangible Cultural Heritage, established in 2003, and its instrumental role in this field. Additionally, we explore the intersection of Big Data visualisation and ICH, discussing how modern data analysis techniques are applied to understand and preserve these invaluable cultural expressions.

START WEBINAR

38:38
Runtime

Ms Kaat De Ridder
Centre for Expertise for Sustainable Business and Digital Innovation, Thomas More University of Applied Sciences, Belgium
Author

b) The role of Big Data in tourism and storytelling

This webinar examines the role and limitations of traditional data in tourism, introduces the concept of Big Data and its advantages in this field and illustrates these points with an example from the UNWTO.

START WEBINAR

24:24
Runtime

Dr Marco Scholtz
Centre for Expertise for Sustainable Business and Digital Innovation, Thomas More University of Applied Sciences, Belgium
Author

c) Big Data analyses for ICH Tourism storytelling

This webinar explores how Big Data analysis can be applied in tourism. We investigate data preparation and analysis, as well as the challenges and potential improvements in data storytelling. Finally, we discuss transforming Big Data insights from Intangible Cultural Heritage (ICH) tourism into compelling narratives.

START WEBINAR

23:44
Runtime

Ms Monika Zdanaviciute
University, Kaunas, Lithuania
Dr Marco Scholtz
Ms Kaat De Ridder
Thomas More University of Applied Sciences
Authors

vii. Webinars for Chapter 7: The future of Big Data storytelling techniques

a) The future of Data Storytelling 1: Introduction to AI-powered Data Storytelling Techniques

Here, we explore the basics of how AI can be used to tell stories from datasets. You will also learn more about how AI works and how it learns human information. We also explore how AI can be trained to interpret situations almost like humans.

START WEBINAR	**16:00** *Runtime*	**Dr Luca Stornaiuolo** *Co-founder Toretei, Milan, Italy* *Author*

b) The future of Data Storytelling 2: The technical aspects of AI-powered Data Storytelling

This webinar examines AI's technical capabilities, which allow it to learn how certain inputs (prompts) influence different information paths that lead to a specific outcome. These paths can be written information or even generated images. Watch this webinar to learn more about how this process works.

START WEBINAR	**18:20** *Runtime*	**Dr Luca Stornaiuolo** *Co-founder Toretei, Milan, Italy* *Author*

c) The future of Data Storytelling 3: Applications and Case Studies

Here, we examine how AI data storytelling can be applied and explore some case studies.

START WEBINAR	**18:33** *Runtime*	**Dr Luca Stornaiuolo** *Co-founder Toretei, Milan, Italy* *Author*

Chapter 1: The Importance of Big Data Storytelling

Mr Gianluca Vagnarelli
Profound Innovation and Social Responsibility
i-strategies, Offida, Italy

TO WATCH THIS CHAPTER VIA WEBINARS, CLICK HERE

This chapter offers a deep dive into the world of Big Data, encapsulating its essence, applications, and transformative impact across various sectors. Starting with a detailed overview of Big Data, the chapter elucidates its fundamental characteristics - volume, velocity, variety, veracity, and value - and how these dimensions form the backbone of modern data-driven strategies. It then ventures into "Big Data Storytelling," exploring the art and science of converting complex, voluminous data into coherent and compelling narratives. This part of the chapter emphasises the crucial role of storytelling in democratising data access and understanding, catering to both technical and non-technical audiences. The narrative progresses to examine the influential role of Big Data in driving social change. This segment highlights Big Data's potential to address significant societal challenges, advocate for social justice, and reshape public policy and activism. Integrating theory with practical insights, the chapter aims to give readers a nuanced perspective on Big Data's multifaceted role in shaping contemporary socio-economic landscapes.

1.1 What is Big Data

The widespread adoption of digital technologies and the extensive array of applications dependent on data has led to the term 'Big Data' becoming common in various fields beyond its original tech-centric domain (see Figure 1.1). This includes disciplines like sociology, medicine, biology, economics, management, and information science, highlighting data's far-reaching impact and relevance across different academic and professional spheres (De Mauro, Greco & Grimaldi, 2016). Originally emerging in the early 2000s within statistics and econometrics, Big Data was used to describe the exponential increase in data quantity and quality due to advancements in data recording and storage technologies (Limaj & Bilali, 2000). This term contains a range of concepts, from massive data aggregation to advanced digital techniques uncovering human behaviour patterns (Favaretto, De Clercq, Schneble, & Elger, 2020). There are various interpretations of what constitutes Big Data. A prevalent view is that Big Data is essentially a dataset characterised by specific traits, such as immense size, rapid expansion, and incompatibility with conventional databases (Chen & Zhang, 2014; Chen et al., 2014; Dumbill, 2013). Another perspective regards Big Data primarily as a technological tool, focusing on the necessary features to achieve specific objectives (Baro et al. 2013). Some studies present Big Data as an amalgamation of data, technology, and analytical methods (Chen et al., 2012; De Mauro et al., 2016; Hashema et al., 2015). Similarly, Boyd and Crawford (2012) perceive Big Data as a multifaceted phenomenon encompassing cultural, technological, and scholarly aspects rooted in the interaction between technology, analysis, and mythology. Sharma et al. (2015) offer a different angle, viewing Big Data as a process undertaken by various entities to achieve outcomes like uncovering new insights. Alternatively, Big Data is seen as an information asset that meets the “V” criteria: volume, velocity, variety, veracity, and value (Beyer & Laney, 2012; De Mauro et al., 2016). That is the definition we are going to explore here.

Figure 1.1: Illustration of the Big Data Society (Source: Created by DALL-E 2, an AI system designed by OpenAI.)

1.1.1 THE FIVE "V" OF BIG DATA

In the context of Big Data, five critical dimensions – Volume, Velocity, Variety, Veracity, and Value – collectively define its complexity and potential. Each of these "Five Vs" represents a unique aspect of Big Data that poses challenges and opportunities for businesses and researchers alike (indicated in Figure 1.2). Let's explore these dimensions in detail:

- **Volume**

Volume is a defining characteristic of big data, referring to the sheer scale of data generated. Every digital process, social media exchange, and connected device contributes to a continuous data stream. This immense volume of data, ranging from petabytes to zettabytes, offers unprecedented opportunities for analysis but also presents significant challenges in data storage, processing, and management.

- **Velocity**

This aspect addresses the incredible speed at which data is generated and must be processed and analysed. In today's digital world, data flows in at an unprecedented rate, necessitating real-time or near-real-time processing to extract timely insights. This rapid data production is driven by 24/7 global connectivity, the Internet of Things (IoT), and users' continuous interaction with digital platforms.

- **Variety**

Big Data encompasses a wide array of data types. This variety includes structured data (organized in databases and spreadsheets), unstructured data (text, images, videos, and social media posts), and semi-structured data (which does not reside in a fixed location but contains tags to organise and process the data). This diversity requires versatile processing tools and methodologies to harness and interpret the data effectively.

- **Veracity**

Veracity refers to the reliability and accuracy of data. In the realm of Big Data, ensuring data quality and accuracy is crucial but challenging, given the varied data sources and types. Veracity affects data decision-making: the higher the quality and accuracy of the data, the more reliable the insights derived from it. Issues like biases, noise, and abnormalities in data must be addressed to maintain the integrity of data analysis.

➢ **Value**

Perhaps the most important "V" value represents the usefulness and actionable insights from processing Big Data. The main objective of analysing Big Data is to extract meaningful and valuable information to aid decision-making, predict trends, and drive innovation. However, the value is not inherent in the data itself; it is realised only when the data is processed and analysed effectively to inform business strategies, scientific research, and policymaking.

Each of these dimensions plays a crucial role in defining Big Data and underlines its complexity and potential in the modern digital landscape. Understanding and addressing these five "Vs" is essential for any organisation or individual looking to harness the power of Big Data effectively.

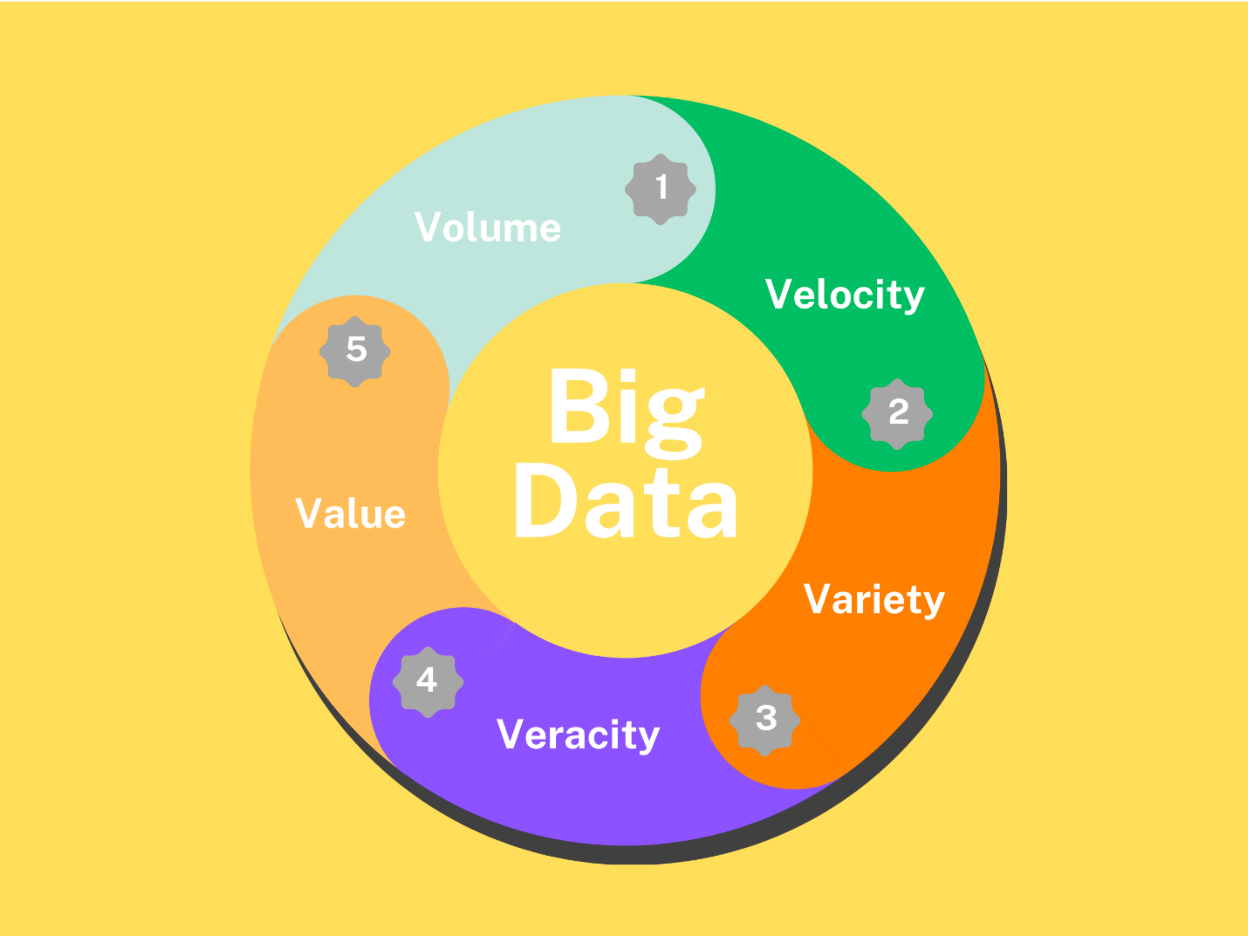

Figure 1.2: Illustration of the five "V" of Big Data definition (Source: Canva)

1.2 Big Data Storytelling

The paragraph on "Big Data Storytelling" delves into the art of transforming complex data into engaging and understandable stories. It discusses the increasing importance of narrative in the era of data overload, where storytelling becomes a crucial tool for making sense of vast amounts of information. This part of the chapter emphasises the power of storytelling in simplifying and giving meaning to Big Data, making it accessible to a broader audience and facilitating informed decision-making.

1.2.1 We live in the narrative era

We live in the narrative era because stories are everywhere (Figure 1.3). Stories are present not only in the traditional field in which stories have always spread, the entertainment industry, but in many other places. Indeed, the application of storytelling traverses a wide range of sectors, from the gaming industry (Carlquist, 2002) to psychology (Kugelmann, 2001), criminology (Presser & Sandberg, 2015), education (Green, 2004), and even the gastronomy sector (Nicolosi & Korthals, 2008). Utilising narrative as a strategic tool, particularly in marketing, is pivotal for communicating the essence of products and brands with more significant impact, as Pulizzi (2012) states. In politics, narratives form an integral part of the communication strategies in political campaigns, weaving through the fabric of candidates' personal stories to connect with the electorate (Gadinger, Mert, Smith & Ochoa, 2016). In the healthcare domain, the individual experiences of patients, narrating their journeys through pain and illness, stand as key components of narrative medicine, a practice brought to the forefront by Charon (2008). As the quoted studies show, stories also permeate areas usually detached from narratives, such as academia, economics, and cultural heritage. Storytelling has been recognised as an engaging method to draw people of varying backgrounds into the participatory process of scientific research, particularly within the scope of Citizen Science (Richter et al., 2019).

Academia has long debated incorporating narrative into scholarly discussions, aiming to complement logical arguments with the narrative's ability to widen audience engagement (Hollihan, Baaske, Riley, 1987). Moreover, the economic discipline, often viewed as analytically stringent, has not been immune to the influence of storytelling. Nobel Prize-winning economist Robert Shiller has posited that economic theories should embrace the role of narratives in shaping ideas, emphasising the human endeavour to create and enhance the contagion of these narratives. He argues that the economy is swayed more by the narratives that capture the public's imagination than by quantitative metrics, with stories spreading through both word of mouth and social media exerting influence on individuals and policymakers alike (Shiller, 2019). Nevertheless, as we mentioned in the beginning, is the entertainment industry the "natural" field of stories, a sector which, particularly in the wake of the "Netflix effect," has seen a substantial embrace of story serialisation (Matrix, 2014).

Figure 1.3: Illustration of the Narrative Era (Source: Created by DALL-E 2, an AI system designed by OpenAI.)

The ubiquitous presence of storytelling across such a broad spectrum of fields signals our entry into an age where narrative is interwoven into every activity. This widespread narrative integration underlines the growing significance of Big Data storytelling. It is a method for simplifying the complex datasets that define our digital era and extracting meaning and insights from them. As we navigate through an ever-expanding ocean of data, Big Data storytelling serves as a lighthouse, guiding us to understand and make better decisions based on the stories our data can tell. In today's digital landscape, where an extensive volume of data is produced, presenting this data intelligibly to the layperson has become crucial, a role that falls to the art of Big Data storytelling (Boldosova & Luoto, 2002).

1.2.2 Big Data Storytelling

In this context, Big Data storytelling emerges as a pivotal tool for making sense of the vast quantities of daily data (see Figure 1.4). This data representation and analysis approach harnesses the power of visualisation and narrative to transform raw data into meaningful insights. Here, we explore five key advantages of Big Data storytelling that demonstrate its effectiveness in conveying information and engaging audiences.

- **Enhance memorability**

Data visualisation can significantly enhance the retention of information by leveraging the human brain's ability to process visual information more efficiently than text. Studies have shown that individuals are more likely to remember visual data than data presented in a text format. This is due to the "pictorial superiority effect," where images are more easily encoded into long-term memory (Nelson, Reed, & Walling, 1976). Furthermore, visual storytelling in presentations has been found to improve recall and comprehension.

- **Engage audience**

Data storytelling engages audiences by transforming numbers and statistics into narratives that resonate on a human level. According to Zak (2014), narratives that induce emotions can lead to greater engagement and empathy from the audience. By weaving data into a compelling story, connecting with the audience emotionally becomes possible, leading to more profound interest and engagement.

- **Increase Big Data understanding.**

In the context of Big Data, where the volume and complexity of information can be overwhelming, data visualisation simplifies the presentation, allowing for easier comprehension. A well-designed visualisation can distil vast datasets into understandable visuals that communicate insights clearly and effectively (Few, 2009).

- **Show patterns**

Data visualisations are particularly adept at revealing patterns and trends that might go unnoticed in textual data presentations. According to Tufte (2001), visualising data can uncover correlations, trends, and outliers that tell a more comprehensive story of the underlying data.

- **Create beauty**

Aesthetic considerations in data visualisation can lead to beautiful representations that convey information and appeal to the viewer's sense of design and artistry. A gorgeous visualisation can capture attention and make the experience of data exploration enjoyable, thus encouraging a deeper interaction with the data (Kirk, 2016).

Figure 1.4: Illustration of the five key advantages of Big Data Storytelling (Source: Canva)

1.2.3 Three examples of data storytelling

We present three practical examples of data storytelling here. The first two ("Description of the Slave Ship" and "La Rutina") are not Big Data storytelling because the data they refer to are small and not digital data. Nevertheless, these two first examples contain aesthetic codes and visual elements crucial to understanding the potential of Big Data storytelling. The third one (Johns Hopkins University's Covid Data) refers to a case of Big Data storytelling.

i. *"Description of the Slave Ship"*

Created in 1787, the "Description of the Slave Ship" diagram (see Figure 1.5) serves as a powerful visual representation of the inhumane conditions aboard slave ships during the transatlantic slave trade. It became an iconic symbol used by social and political movements advocating for the abolition of the slave trade. Its widespread appearance in newspapers, pamphlets, books, and posters made it highly effective in raising public awareness about the brutal realities of slavery.

- **Visualising the Harsh Reality**

The diagram graphically depicts the cramped and deplorable conditions in which enslaved Africans were transported. It illustrates how each person was confined to a tiny space, often chained together, with inadequate food and water, leading to the spread of diseases and many deaths during the voyages.

- **Effective Data Representation**

This image is notable for effectively illustrating data about the slave trade that had never been visualised so powerfully. Its immediate impact of horror upon viewers highlights the importance of how data is presented in evoking public response and fostering social change.

- **Aesthetic Codes in Data Storytelling**

  - *Non-fictional Approach*: The diagram depicts a real slave ship, emphasising that the individuals shown are real people, not abstract figures. This realism makes the data representation more impactful.
  - *Counternarrative*: By providing a raw and factual depiction of slavery, the diagram challenges and dismantles idyllic narratives about the slave trade, offering a more truthful account.

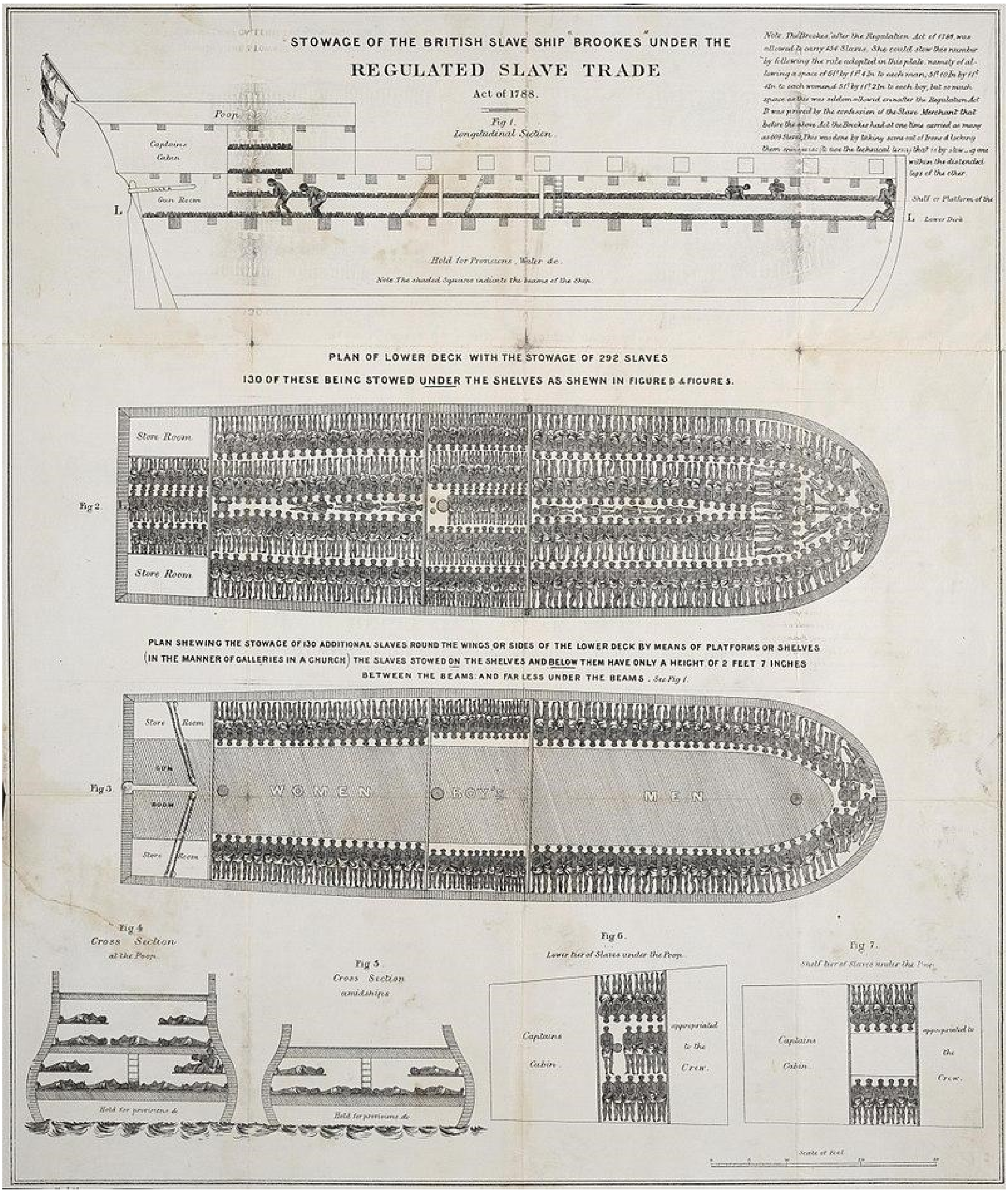

Figure 1.5: A plan of the British slave ship Brookes, showing how 454 enslaved people were accommodated on board after the Slave Trade Act of 1788. Published by the Society for Effecting the Abolition of the Slave Trade (Source: [Wikipedia])

- *Empathy Induction*: The portrayal of enslaved individuals in a passive and victimised state evokes empathy in viewers, effectively conveying the sense of suffering and claustrophobia experienced by the enslaved people.

- *Evocative Imagery*: The diagram resonates with cultural and biblical imagery, particularly the story of Noah's Ark, yet presents a stark contrast of disorder, greed, and fear, enhancing its memorability.
- *Semiotic Shock Tactic*: The traditional style of the diagram, combined with the stark depiction of human bodies, creates a semiotic shock that underscores the gravity of the slave trade. This human aspect of the diagram makes it a powerful tool in the representation of data and a catalyst for social change.

In summary, “A Description of the Slave Ship” is a pivotal example of data storytelling, demonstrating the profound impact visual representation of data can have in shaping public opinion and advancing social causes. The diagram not only presents data but also tells a story that connects with viewers on a deeply human level, emphasising the crucial role of storytelling in the realm of data and social change.

ii. "La Rutina" - A Story of Love and Flowers

The project, titled "La Rutina" (meaning "Routine" in English), was created by Spanish information designer Jaime Serra. It presents a unique visual representation of a man giving flowers to his love over 48 years. This project beautifully evocates their love story through data visualisation.

- **A love story behind the data**

This love story began in 1962 in Spain. A 22-year-old peasant met his 16-year-old love and presented her with a yellow flower he had picked from the fields. Marking the beginning of their journey, she preserved this flower. This tradition continued for forty-eight years, encompassing three years of courtship and forty-five years of marriage, until the man's passing.

- **Visualising the journey of love**

In "La Rutina," each flowerpot depicted in the visualisation represents the number of flowers given yearly (Figure 1.6). Some years are abundant, while others are sparse. Interestingly, the flower colours change from yellow to red, signifying the couple's move from the countryside to the city, where the man can no longer find the same yellow flowers. One year features an empty pot, reflecting a time when he was too ill to pick a flower. The narrative culminates in 2010 with another empty pot, marking the end of the love story.

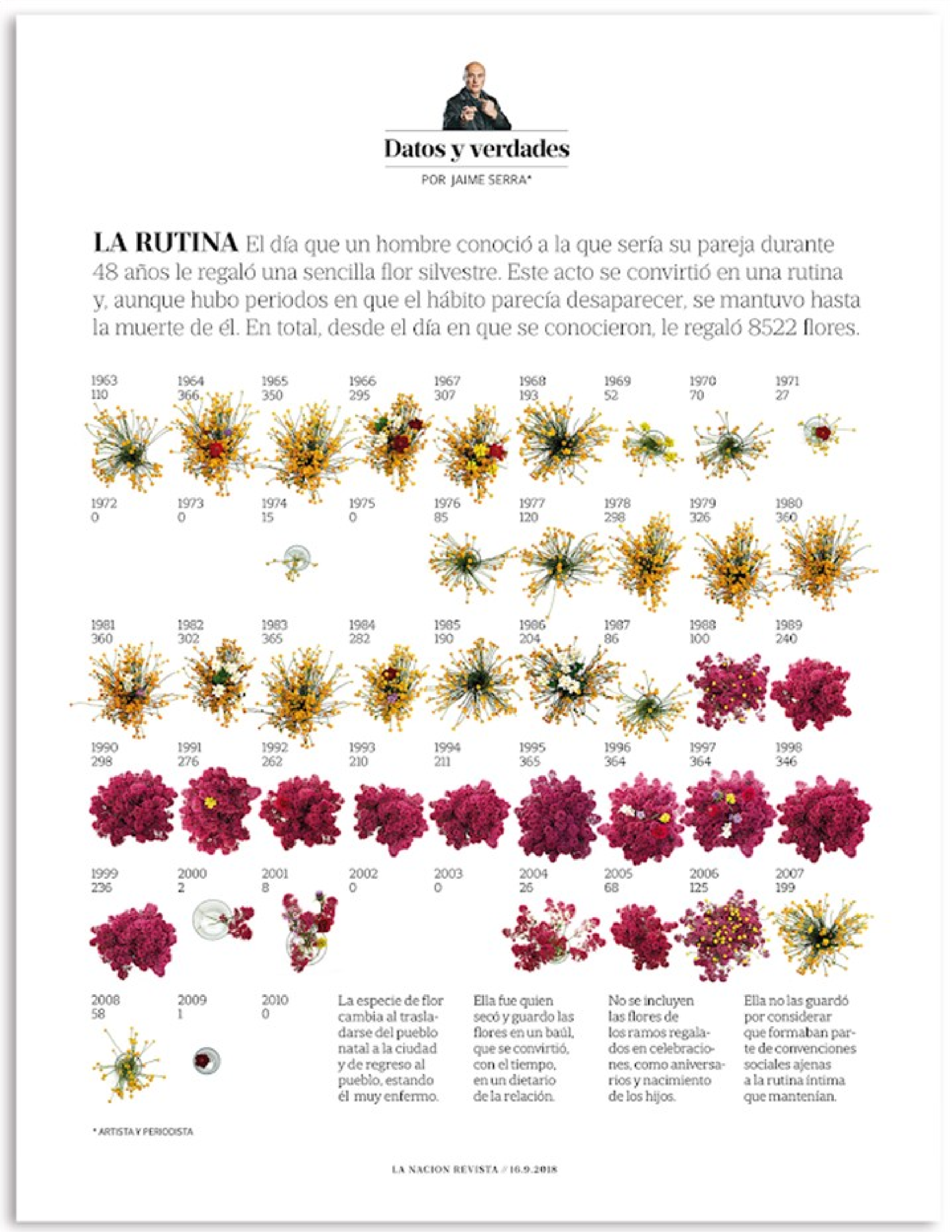

Figure 1.6: La Rutina by Jaime Serra (Source: https://jaimeserra-archivos.blogspot.com/search?q=rutina).

- **Key Insights from "La Rutina"**

This project illustrates two critical aspects of data storytelling:

i. ***Blending 'Cold' Data with a 'Warm' Story***: The project masterfully combines quantitative data with the emotional depth of a love story. This blend highlights how storytelling can enhance the appeal of data. The key takeaway is the importance of the human element in data representation. In "La Rutina," the love story between the man and woman breathes life into the data, creating an emotional connection with the audience.

ii. ***Narratives from Everyday Data***: "La Rutina" demonstrates that engaging narratives can be crafted from everyday data. This insight is precious for educational purposes, encouraging students to develop their data storytelling skills using accessible data sets. This approach especially benefits non-ICT students lacking technical skills in handling Big Data. The project underscores creativity and narrative crafting as essential competencies for compelling data storytelling, regardless of the data's scale.

In conclusion, "La Rutina" exemplifies how data storytelling can transcend the boundaries of Big Data, using small-scale data to weave narratives that resonate deeply with audiences. This project serves as a testament to the power of storytelling in bringing data to life and engaging viewers on a human level.

iii. Johns Hopkins University's Covid Data

During the COVID-19 crisis, many governments, including Italy, actively disseminated data to the public, offering regular updates on new infections, casualties, and other vital statistics. The Johns Hopkins University's Coronavirus Resource Center was a critical global resource for pandemic data (example in Figure 1.7). This platform collated open data on COVID-19 from countries worldwide, allowing for a comprehensive, global view of the pandemic's status and enabling comparisons between nations.

➢ **Empowering Public Opinion and Civil Society**

This widespread sharing of Big Data was crucial in informing the public about the pandemic's evolution. But more importantly, it empowered citizens and civil societies to engage with the data beyond the interpretations provided by public authorities. Access to this data meant individuals could analyse, interpret, and draw conclusions independently, leading to informed public debates and discussions.

➢ **Case Study: Italy's Public Debate**

A notable example of this was observed in Italy. The availability of COVID-19 data led to public debates concerning the government's response and strategies to combat the virus. These debates were fuelled by diverse interpretations of the same data sets released by the government. Citizens with access to data and analytical tools could question, support, or critique governmental policies based on their understanding of the pandemic's data.

This phenomenon underscored a significant shift in how public health crises are managed and discussed in the information age. The availability of Big Data not only informed but also democratised the discourse surrounding COVID-19, allowing a broader section of society to participate actively in critical discussions about public health strategies and decisions.

In summary, the case of data sharing during the COVID-19 pandemic illustrates the powerful impact of Big Data in shaping public opinion and enabling civil society to engage in meaningful discourse based on factual and analytical understanding. It highlights the importance of data accessibility and transparency in fostering informed public debates and empowering citizens in decision-making processes.

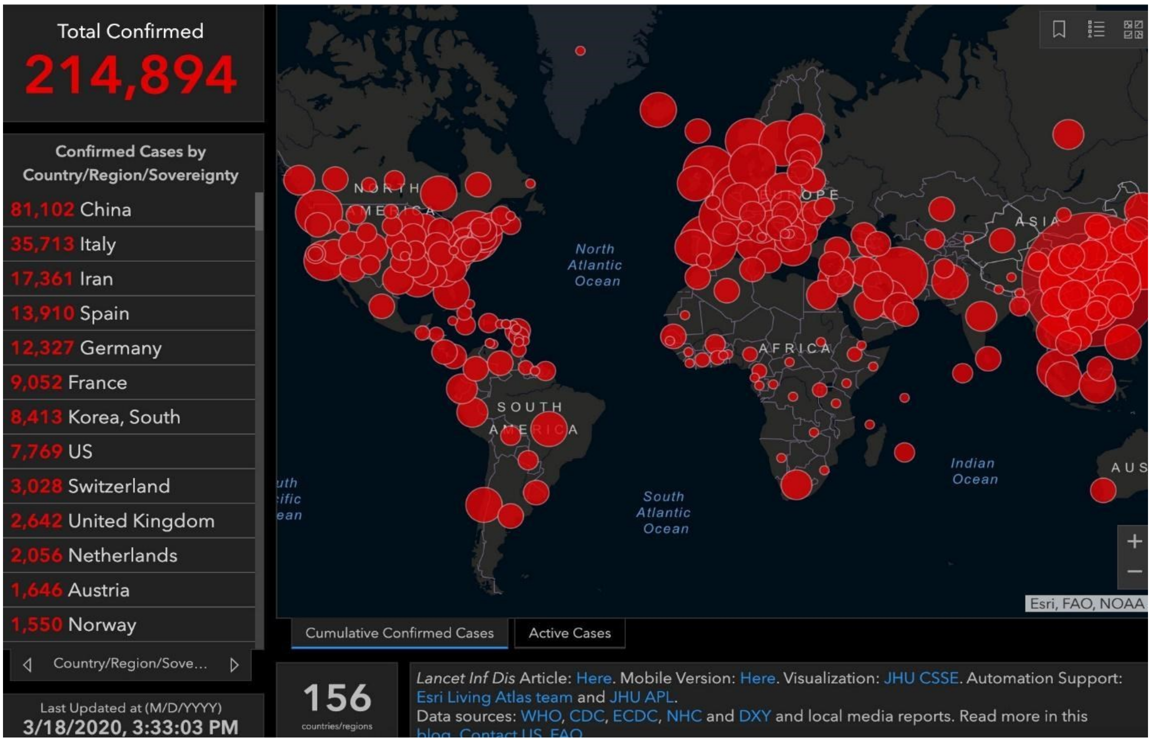

Figure 1.7: Johns Hopkins University's Coronavirus Resource Center Pandemic Data (Source: https://coronavirus.jhu.edu/map.html).

1.3 Big Data for Social Change

The "Big Data for Social Change" paragraph focuses on applying big data storytelling to address and reshape societal challenges. It highlights how Big Data transcends traditional analytics, becoming a tool for promoting equity, justice, and responsive social environments. This part of the chapter discusses the role of Big Data storytelling in activism, philanthropy, and social justice, illustrating its potential to drive significant and meaningful societal changes.

1.3.1 Big Data for Social Change Meaning

In an age where the volume, velocity, and variety of data are expanding exponentially, the concept of Big Data for social change emerges as a transformative force in addressing and reshaping societal challenges (see Figure 1.8). This concept transcends the traditional boundaries of data usage, leveraging the power of large datasets to foster equitable, just, and responsive social environments. Big Data storytelling and visualisation are central to this paradigm shift, which plays critical roles in translating complex datasets into narratives and visuals that are accessible and compelling to a broad audience. When presenting concrete examples of Big Data storytelling for social change, it is necessary to clarify what Big Data for social change means. We refer here to four meanings of this definition:

- **Data Activism**

This form of Big Data for social change involves individuals and organisations actively employing data to drive social and political change. Data activism can take various forms, such as utilising data to raise awareness about social issues, influence policy, or challenge existing power structures. It encompasses proactive measures (like leveraging open data to instigate societal reforms) and reactive approaches (such as safeguarding privacy and resisting surveillance). Data activists analyse, interpret, and deploy data to reveal insights and narratives that might otherwise remain obscured, empowering citizens and communities (Williams, 2020).

- **Data Philanthropy**

Data philanthropy represents the practice of private sector organisations donating their data resources for the public good. It involves companies sharing their vast data pools to aid humanitarian causes, public health initiatives, and environmental protection efforts. This approach to Big Data allows for harnessing typically inaccessible corporate data for societal benefit, exemplified by initiatives like the Data for Development (D4D) Challenge. Data philanthropy reflects a growing awareness in the private sector of the societal value of data and the importance of contributing to the common good.

- **Counter Mapping**

Counter-mapping, or counter-cartography, is the process of creating maps that challenge conventional or mainstream geographical representations. It serves as a tool for marginalised communities to visualise and narrate their experiences, histories, and connections to the land. Counter-mapping can expose social injustices, environmental threats, and cultural erasures that mainstream maps may overlook or omit. This form of Big Data for social change application empowers communities to assert their rights, influence policy decisions, and reshape public perceptions about space and place.

- **Data Justice**

Data justice is a relatively new field that focuses on the ethical implications of data collection and use, emphasising fairness, equity, and inclusion in how data is gathered, analysed, and applied. It addresses concerns about bias, discrimination, and inequality in extensive data practices, advocating for the rights of individuals and communities in the digital age. Data justice seeks to ensure that Big Data technologies do not reinforce existing social injustices but are used to promote more equitable and just outcomes. It involves scrutinising data policies, practices, and algorithms to safeguard against the perpetuation of systemic biases and to advocate for the fair treatment of all people, regardless of their background.

Figure 1.8: Illustration of the four types of Big Data for Social Change (Source: Canva)

These four types of Big Data for social change (Figure 1.8) – data activism, data philanthropy, counter-mapping, and data justice – highlight how data can be leveraged to foster a more equitable, just, and responsive society. Each approach offers unique strategies and perspectives for using Big Data to address societal challenges, underscoring the potential of data as a powerful tool for social good.

1.3.2 Practical Examples of Big Data for Social Change

i. Periscopic gun killings in the U.S. (www.guns.periscopic.com)

One exemplary instance of Big Data storytelling for social change is the "Guns Periscopic" project, which addresses the issue of gun violence in the United States. This project aims to raise awareness about the impact of gun violence by illustrating the years of life lost due to such incidents. As users visit the page, they are greeted with the ongoing count of gun-related fatalities for the year 2018, alongside a statistical estimation of the potential years these individuals could have lived.

Big Data storytelling in this context encompasses three key elements:

- **Shape as a Narrative Tool**

The project utilises an arc shape to represent gun violence. This arc symbolises life's journey, encapsulating stages from birth to natural death. The arc shape is intentional, designed to evoke the passage of time and the evolution of life. It serves as a reminder that shapes in Big Data storytelling are not arbitrary but are deliberately chosen to convey specific narratives.

- **Colour as a Communicator**

The project uses specific colours to differentiate lived life from potential life lost. The lived part of the arc is depicted in warm colours like orange and yellow, commonly associated with energy, passion, and life, reflecting the time the victims were alive. In contrast, the lost life is illustrated in grey against a dark background, colours often associated with mourning and sadness in Western cultures. This choice of colours is purposeful, enhancing the storytelling by aligning with the conveyed message.

- **Text for Personalisation**

The project incorporates short texts alongside the visual elements of shape and colour. These texts bring a personal dimension to the data, sharing individual stories of gun violence victims. This aspect is crucial in Big Data Storytelling, as it allows for emotional connection with the audience, moving beyond mere statistics to humanise the data. The text complements the visual elements, providing a deeper, more empathetic understanding of the data.

In summary, the “Guns Periscopic” project transforms the usual data presentation on gun violence, which often focuses solely on victim counts. Instead, it shifts the narrative to the potential years of life lost due to such violence, offering not just an informative representation but an emotionally resonant and counterfactual visualisation. The use of grey colour visually signifies the life that could have been lived if not cut short by violence.

The overall effect of this project is a poignant sense of nostalgia for a future that was denied by violence. It is a unique feeling of sorrow for what could have been, encouraging empathy for the victims. This project is an insightful and innovative approach to Big Data storytelling, demonstrating how data can be used to tell powerful, emotive stories that resonate with audiences.

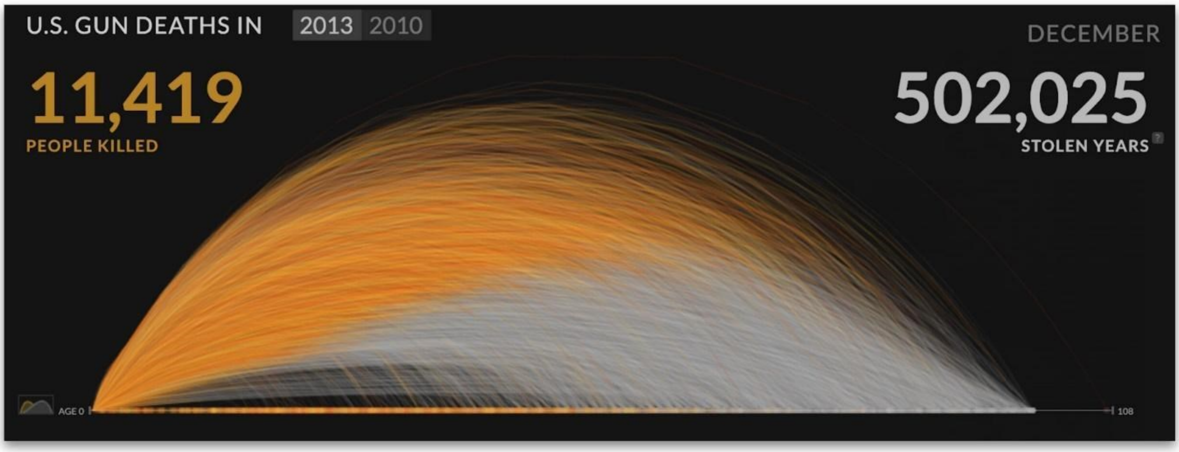

Figure 1.9: Guns Periscope Project (Source: www.guns.periscopic.com)

ii. Mapping Diversity (www.mappingdiversity.eu)

Critical Cartography, as a concept, challenges traditional mapping practices by revealing the biases and power structures inherent in conventional cartography. Coined as "mapping the unmapped" by the MIT SLAB in 2017 and described as counter-mapping by Peluso (1995), it posits that maps, much like written texts, images, or films, are not value-neutral. Instead, they reflect and perpetuate power dynamics, as Perkins (2018) noted. This approach to cartography underscores the idea that maps are not just geographical representations but also political and cultural narratives.

A compelling example of this is the "Mapping Diversity" project, which also embodies principles of data feminism. This initiative is a platform designed to explore and discuss diversity and representation in European street names. Their website states the project analysed 145,933 streets in 30 major European cities across 17 countries. The startling finding was that over 90% of streets named after individuals honour white men, raising critical questions about the visibility and representation of other groups in Europe's urban spaces.

The "Mapping Diversity" website allows users to interact with the data, offering insights into each city involved in the study. Users can compare the diversity scores of European cities, identify the most popular female figures in street names, and gauge the inclusiveness of different locales. This project stands apart in its aesthetic approach to data presentation, highlighting the underrepresentation of women in street toponyms. By comparing data across cities, the project also makes the varying levels of inclusiveness regarding diversity apparent.

Figure 1.10: Mapping Diversity Project (Source: www.mappingdiversity.eu)

"Mapping Diversity" exemplifies a new form of social practice in critical cartography, where data is conceptualised through political activism. The primary aim is to raise public awareness about the unequal representation in European city names. This project embodies Big Data for social change, utilising data to critique reality and create unconventional narratives that mobilise public opinion around pressing social issues.

In summary, the "Mapping Diversity" project showcases how Big Data can be leveraged for analysis and as a catalyst for social change (Figure 1.10). It emphasises the importance of representation in our urban spaces and challenges us to rethink the narratives that our cities embody and perpetuate.

1.4 The European Strategy for Data

As the importance of Data and Big Data grows, it is also important that governing agencies take this into account to ensure that Big Data's power is harnessed in a sustainable manner, ensuring access and privacy at the same time. To learn more about this, watch this webinar by Mr Alessio Ceci from i-Strategies in Italy.

1.5 Conclusion

In conclusion, the chapter reaffirms Big Data's potential and growing significance in the modern world. It has highlighted the technical aspects of Big Data and its human-centric applications, particularly in storytelling and social change. By distilling complex data into engaging narratives, Big Data storytelling emerges as a vital tool in bridging the gap between data scientists and the public, fostering a more inclusive understanding of data insights. Furthermore, the chapter emphasises the pivotal role of Big Data in driving social change. It illustrates how data-driven approaches can provide novel insights into societal issues, offering a new lens to view and address global challenges. The conclusion summarises the essence of Big Data as a catalyst for innovation and change, driving forward advancements in technology, society, and policy. This chapter, therefore, serves as a vital resource for understanding the profound impact of Big Data across various domains, inspiring readers to consider the endless possibilities of data-driven decision-making and action.

Chapter 2: The Unreasonable Effectiveness of Big Data

Prof. Haibo Li
Department of Media Technology and Interaction Design
KTH, Royal Institute of Technology, Sweden

TO WATCH THIS CHAPTER VIA WEBINARS, CLICK HERE

Big Data technology has gained significant popularity, and its application to real-world problems is increasingly widespread. Despite this, the effectiveness of Big Data technology is not yet fully understood. This presents a technical paradox: the tension between computational complexity, often known as the 'curse of dimensionality', and the proven exceptional effectiveness of Big Data solutions. This chapter provides a mathematical perspective to enhance our understanding of how these data analytics systems work, evaluate the strengths and weaknesses of various model architectures and potentially lead to significant advancements. We believe that the unreasonable effectiveness of Big Data can give us unexpected benefits, revolutionising how we approach problem-solving and decision-making.

2.1 INTRODUCTION

Big Data technology has become a cornerstone in the business and scientific worlds, driven by advancements in data analytics techniques, storage solutions, and processing power. Big Data tools, such as machine learning algorithms, statistical models, and data mining techniques, can be used to explore and identify hidden associations in large and complex datasets. The hidden association is also called an "unexpected association, " which refers to discovering surprising or non-obvious relationships between different variables or data points. Identifying such associations can drive innovation, offering significant benefits in various areas. For example, in marketing, it can help discover new customer segments; in healthcare, it aids in identifying previously unrecognised disease risk factors; and in finance, it assists in uncovering complex market trends.

While Big Data analytics offers valuable insights, it requires careful handling to ensure validity and ethical integrity. For example, one of the challenges with these associations is ensuring they are meaningful and not just random patterns. It requires careful analysis and often domain expertise to distinguish between genuinely significant relationships and spurious correlations. In this chapter, we offer a mathematical perspective on several challenging but interconnected questions: *Why do unexpected associations arise*? *What does the term 'unreasonable* *effectiveness' imply*? *Why is this effectiveness considered unreasonable*? And *why do spurious correlations emerge*?

2.2 The Effectiveness of Big Data

The effectiveness of Big Data is evident in at least two key aspects:

- **Unexpected Patterns and Correlations:**

Big Data analysis can reveal patterns and correlations that are not immediately apparent. These may be relationships between different data elements that were not previously known or considered.

- **More Data, Better Performance:**

In computer vision, continual improvements have been observed due to using larger volumes of visual data to train deep neural networks.

2.2.1 The Unexpected Association

A popular story frequently mentioned in discussions about data analytics goes as follows: Once upon a time, in a supermarket, the management decided to dive deeper into their sales data to understand customer behaviour better. They hired a data scientist to work on the data to find patterns and associations in their customers' purchasing habits. As they sifted through mountains of transaction data, a surprising association emerged: many customers who bought diapers also bought beer. This correlation was unexpected and, at first glance, somewhat humorous. Why would these two seemingly unrelated items be frequently purchased together?

Intrigued, the store's management decided to investigate further. This pattern often occurred during certain times, mainly late afternoons and evenings. Diving into the context, they learned that fathers often made these purchases. The typical scenario was that these fathers, tasked with buying diapers for their young children, were also picking up beer for themselves.

Seeing an opportunity, the store experimented by placing diapers and beer closer together in the store layout. To their delight, they observed a noticeable increase in the sales of both products. This strategic change directly resulted from the insights gained from data analytics.

The story of beer and diapers became a legend in retail, highlighting how data analytics can reveal hidden patterns that lead to innovative business strategies. It showed that even the most unlikely product pairings could have a solid underlying rationale waiting to be discovered through data analysis.

But one might wonder why the associations discovered are often labelled' unexpected.' This is primarily because they frequently involve unlikely pairings. The reason behind this is not the volume of the data per se. Instead, it can be attributed to the combinatorial explosion,' a mathematical term that describes the rapid growth in problem complexity due to increased parameters, posing significant computational and analytical challenges. For instance, if we consider N items, they typically could form $l = \frac{N(N-1)}{2} = \binom{N}{2}$ pairs. With a thousand items (where N=1000), the total possible combinatorial pairs amount to about half a million (500,000 items), which naturally includes many unexpected associations, often extending beyond one's imagination. Here, it's important to note that while this growth is often casually referred to as 'exponential,' it is actually polynomial. However, when we consider combinations of more than two items, we encounter a combinatorial explosion. The number of all possible combinations becomes $2^N - 1$, where N is the number of items. Managing such an immense number of combinations is an exceedingly challenging task!

2.3 Effectiveness in Vision Tasks

Over the past decade, computer vision has witnessed remarkable advancements, a significant portion of which is attributable to the application of deep learning models. A key question has been posed at Google: "*If we increase the training data tenfold, will the accuracy double? What if we amplify it a hundredfold or even three hundredfold? Will the accuracy plateau, or will we observe continual improvements with the addition of more data?* " (Gupta, 2017). In pursuit of answers, Google has dedicated efforts toward constructing such datasets automatically to refine computer vision algorithms. In their recent paper, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era", they observed a logarithmic-like relationship, where the mean average precision (meanAP) is expressed as $meanAP = k\log_{10}N$; here, k represents a constant, and N denotes the unit of data volume. This relationship illustrates the correlation between performance in vision tasks and the amount of training data utilised in representation learning, as illustrated in Figure 2.1. Their results underscore the extraordinary impact of data quantity on performance enhancement, affirming the principle 'more data, better performance.' Other researchers have confirmed this observation, explaining the extensive collective efforts invested in data acquisition and management.

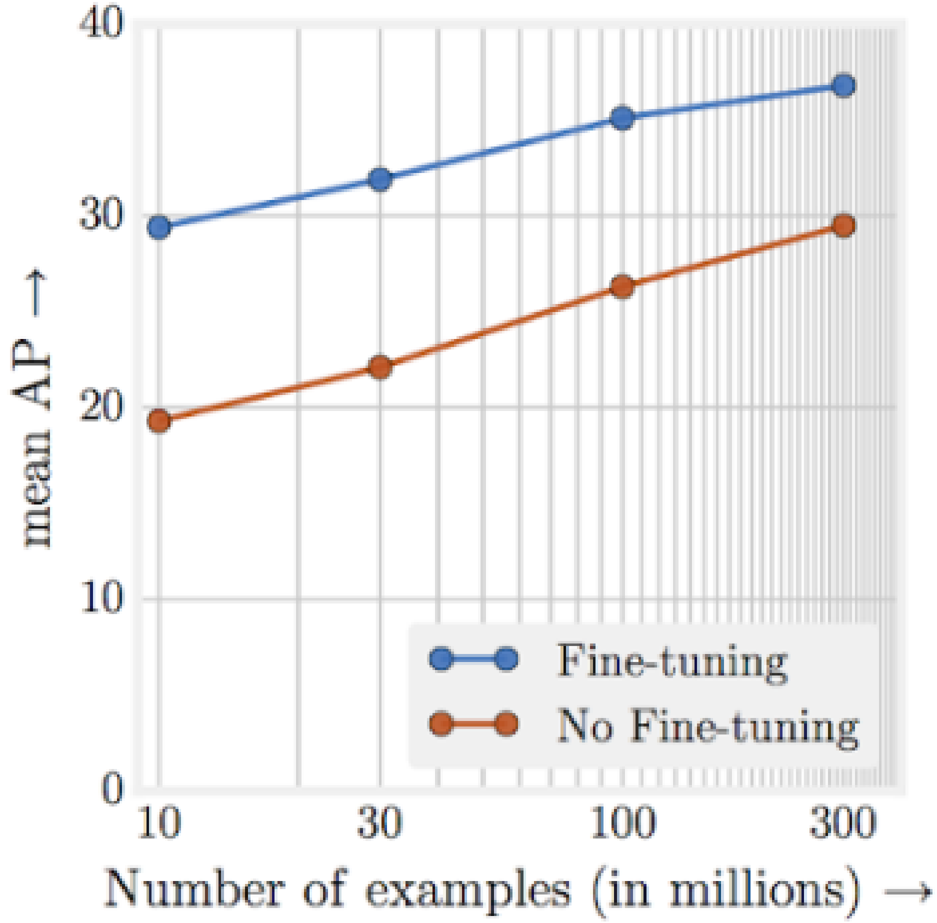

Figure 2.1: Object detection performance when pre-trained on different subsets of JFT-300M from scratch. The X-axis is the dataset size in the log scale, and the y-axis is the detection performance in mAP@[.5,.95] on the COCO-minimal subset (Wiegner, 1960).

2.4 A Mathematical Explanation of The Unreasonable Effectiveness of Big Data

The effectiveness of Big Data lies in the combinatorial interactions among N items, typically scaling at an order of $O(2^N)$. Even when considering just two items, the complexity reduces to a polynomial order of $O(N^2)$, which still offers a vast array of possible pairings and leads to unexpected combinations. However, this immense complexity also poses significant, intractable challenges to Big Data techniques, creating a well-known paradox in data analytics. The paradox often challenges our understanding of Big Data and prompts deeper investigation or rethinking of accepted principles.

In practice, effective Big Data algorithms can address these challenges. The Apriori algorithm is a fundamental tool in data analytics for discovering frequent item sets and deriving association rules, helping analyse complex patterns in large datasets. While these algorithms are practical, the question arises: why do we refer to this as ’unreasonable effectiveness’? What exactly is ’unreasonable’ in this context?

The term ”unreasonable” generally refers to something that is not guided by or based on good sense, logic, or clear thinking. It has been used in various contexts to describe actions, decisions, expectations, or behaviours that are irrational, excessive, or unjustified by the circumstances. When we refer to Big Data techniques as possessing ’unreasonable effectiveness’, we acknowledge that their success or utility defies our intuitive understanding, everyday experiences, and logical reasoning. Intuitively, we might expect that the immense complexity inherent in Big Data would render these techniques intractable. However, existing Big Data methods have proven surprisingly effective in uncovering hidden associations and excelling in data modelling and analytics.

In the following sections, we will provide a mathematical explanation for the seemingly unreasonable effectiveness of Big Data techniques.

2.4.1 Birthday Problem

To understand the unreasonable effectiveness, let us first introduce the famous birthday problem in probability theory (Gardner, 1976). The birthday problem asks for the probability that, *in a set of **n** randomly chosen people, at least two will share a birthday*. The birthday paradox refers to the counterintuitive fact that only 23 people are needed for that probability to exceed 50%.

The probability $P(A)$ that at least two people in a room share the same birthday can be more easily calculated indirectly. First, we calculate $P(A')$, the probability that no two people in the room have the same birthday. Then, because A and A' are the only two possibilities and are also mutually exclusive, we have $P(A) = 1 - P(A')$.

$$P(A') = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \cdots \times \frac{343}{365}$$

The terms of the equation can be collected from the example to arrive at the following:

$$P(A') = \left(\frac{1}{365}\right)^{23} \times (365 \times 364 \times 363 \times \cdots \times 343)$$

Evaluating the equation gives $P(A') \approx 0.492703$. Therefore, $P(A) \approx 1 - 0.492703 = 0.507297$.

The birthday problem is a paradox: it seems wrong at first glance but is, in fact, true. While it may seem surprising that only 23 individuals are required to reach a 50% probability of a shared birthday, this result is made more intuitive by considering that the birthday comparisons will be made between every possible pair of individuals. With 23 individuals, there are pairs to consider, far more than half the number of days in a year.

Let us see how to compute the probability approximately. A good rule of thumb which can be used for mental calculation is the relation

$$p(n) \approx \frac{n^2}{2m} \qquad (1),$$

where n is the number of people in the room, and m is the number of days in a year (all birthdays). The relation can also be written as

$$n \approx \sqrt{2m \times p(n)} \qquad (2),$$

which works well for probabilities less than or equal to ½. In these equations, *m* is the number of days in a year.

For instance, to estimate the number of people required for a $\frac{1}{2}$ chance of a shared birthday, we get

$$n \approx \sqrt{2 \times 365 \times \frac{1}{2}} = \sqrt{365} \approx 19 \qquad (3)$$

which is not too far from the correct answer of 23.

The birthday paradox exemplifies data volume's remarkable efficacy in probing a problem's entire space. If the problem space is represented as M, then a size dataset $n = \sqrt{M}$ is adequate to tackle the problem effectively.

Revisiting our previous example, if we have a thousand items, with $N = 1000$, the number of potential combinations reaches as high as $2^N - 1$. Clearly, the vast majority of these combinations are random and lack any meaningful significance. In the context of a supermarket, any transaction made by a customer represents the realisation of these 'meaningful' combinations of items. The number of these significant combinations is considerably smaller; for instance, let's say one million, denoted as $M = 10^6$. The challenge is that these *M* meaningful combinations are unknown. The primary objective of data analytics is to uncover these hidden combinations. The birthday problem suggests that the same combination of items can be found in just 1000 (≈$\sqrt{M} = \sqrt{10^6}$) transactions. As a result, analysing just 1,000 transactions is sufficient instead 4 of waiting for a million transactions to begin making sense of the data. This approach is 1000 times more effective, representing an increase of effectiveness by three orders of magnitude. The counterintuitive nature of the birthday problem illustrates the unreasonable effectiveness of Big Data.

2.4.2 Same Birthday as Yours

Now, let's delve into why more data leads to better performance in the realm of Big Data. To illustrate this concept, we can introduce a variant version of the birthday problem.

In the birthday problem, neither of the two people is chosen in advance. By contrast, the probability $q(n)$ that someone in a room of n other people has the same birthday as a particular person (for example, you) is given by

$$q(n) = 1 - \left(\frac{365 - 1}{365}\right)^n \tag{4}$$

and for general d by

$$q(n; d) = 1 - \left(\frac{d - 1}{d}\right)^n. \tag{5}$$

In the standard case of d = 365, substituting n = 23 gives about 6.1%, less than one chance in 16. For a greater than 50% chance that one person in a roomful of n people has the same birthday as you, n would need to be at least 253. This number is significantly higher than $\frac{365}{2}$ = 182.5. The reason is that it is likely that there are some birthday matches among the other people in the room. This type of problem is referred to as the '*same birthday as yours*' problem.

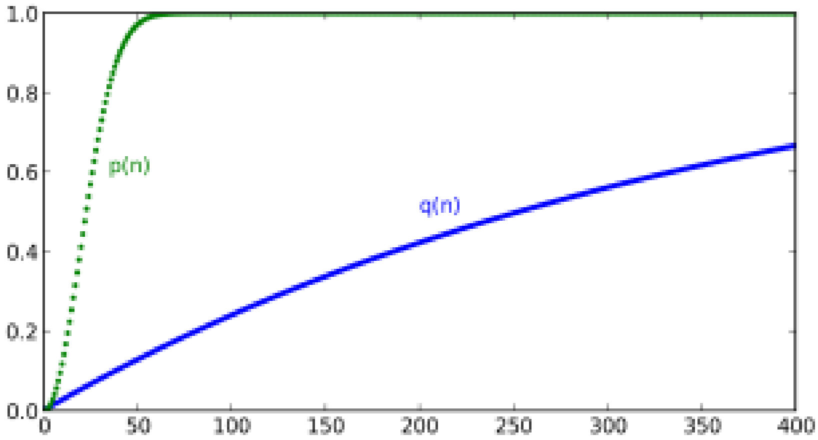

Figure 2.2: The computed probability of at least two people sharing a birthday versus the number of people. Comparing p(n) = probability of a birthday match with q(n) = probability of matching your birthday.[2]

Now, we illustrate how to map an object detection task in computer vision to the birthday problem. We assume that a given deep network model possesses a specific capacity for discrimination, which is helpful for object detection. This capacity can be quantified by entropy or by assuming the model could classify maximally **d** distinct classes. When training the model with a dataset of **n** images, each containing only one object. Without loss of generality, we further assume that the model is well-trained and can accurately classify the **n** images into the correct classes **c**. Here, the number of classes **c** should be less than or equal to the model's capacity **d**.

If we consider the object labels as 'birthdays', then the model's capacity, **d**, represents the number of days in a 'year'. Given an input image, the network generates its 'birthday'.

For an object in a given image to be successfully detected, the training dataset must contain at least one image containing the same object or an object with an identical 'birthday'. This transforms the object detection problem into determining the probability of encountering another image containing an object with the same 'birthday'. The larger the number of images in the training dataset, the higher the probability that an object in a given image will encounter another object in the dataset with an identical 'birthday'. From the 'Same Birthday as Yours' problem, we know this probability is given by equation (2), and it is directly proportional to the Mean Average Precision (mAP), a key performance metric in object detection. The relationship between the probability and the data size **n** is depicted in Figure 2.3, where we observe an approximate logarithmic relationship between the probability $q(n;\, d)$ and the size of training data n used in representation learning. The equation delineates this relationship.

$$q(n; d) = 1 - \left(\frac{d-1}{d}\right)^n .$$

As n increases, the probability of a match in the set grows in a manner that resembles a logarithmic function of the volume of data. This observation explains the (unreasonable) effectiveness of data in this context, as claimed by Google.

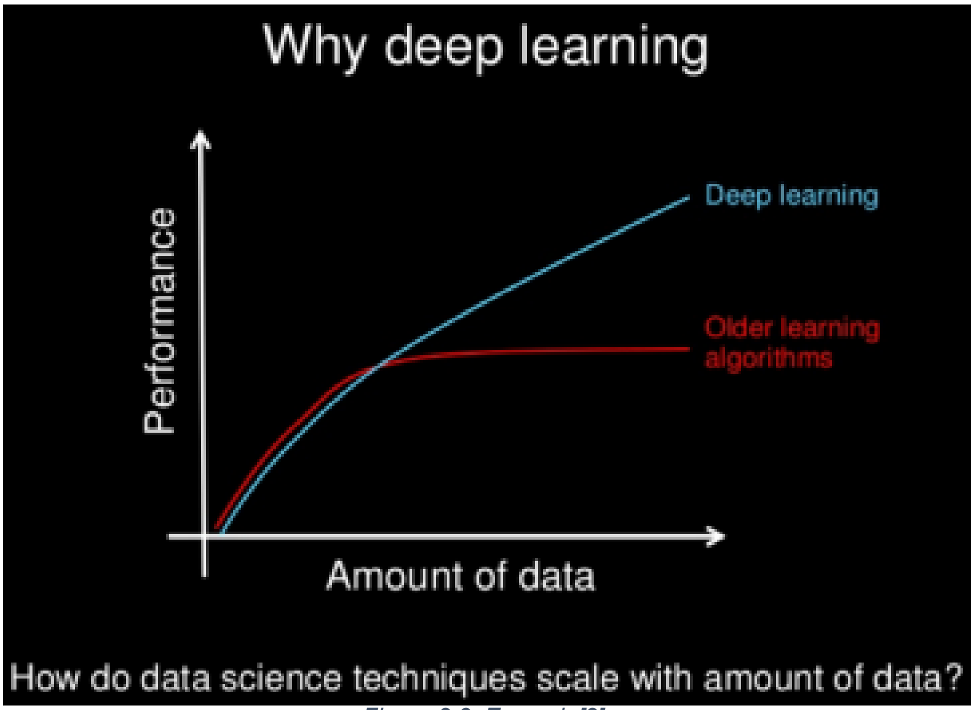

Figure 2.3: Example[2]

Big Data's unreasonable effectiveness' can only be illustrated in large neural networks. In data science, the intuitive belief that observing more data leads to better performance holds only within a specific range. When the data size reaches a certain limit, performance tends to plateau. Counterintuitively, with more extensive or deeper networks, more data seems to enhance performance consistently, as illustrated in Figure 2.3. The capacity of the network d can explain this.

Large datasets hold immense information about underlying patterns, proving highly effective in facilitating feature learning. This significantly aids large models in accurately assigning the correct birthday labels to individual objects. For example, a dataset of millions of images might contain subtle variations in lighting, colour, and texture that are difficult for a human expert to identify but can be learned by a machine learning algorithm when large models are employed. By training on such a large and diverse dataset with an extensive network model, the algorithm can learn to generalise the assignment of birthday labels to new examples much more effectively than if it had only seen a small subset of the data.

2.5 Combinational Exploration and Spurious Relation

The underlying mechanism for the unreasonable effectiveness of Big Data lies in the phenomenon of combinatorial explosion. In a given set containing *N* neural network attributes, approximately 2^N hypotheses could be generated. This vast array of hypotheses presents a ground for uncovering unexpected matches or associations. However, recognising that we often discover correlations, not causations, is crucial. For instance, the increased beer sales are not caused by the sales of diapers. There is no direct cause-and-effect relationship between these two. The actual underlying factor, also known as 'confounder', is the behaviour of young fathers who purchase beer as a personal reward on Friday afternoons, typically after buying diapers. A confounder can influence both of the variables being analysed. This confounding variable creates the illusion that the two primary variables are directly related to each other.

We must be careful, as the most unexpected matches often correspond to spurious relationships. This is due to each pair of combined attributes representing a potential hypothesis. The challenge arises from the sheer number of hypotheses, which are so vast that there is seldom enough data to discern which hypotheses are true and which are false accurately. This situation typically leads to an overfitting problem, a well-known and troublesome issue in machine learning. Overfitting tends to result in spurious relationships, which can lead to profound ethical implications.

2.6 Concluding Remarks

The term 'unreasonable effectiveness of Big Data data' highlights the extraordinary capacity of large datasets to enable effective learning and yield highly accurate models, even using relatively simple algorithms. This is possible due to the vast amount of information in large and diverse datasets, which machine learning algorithms can learn to generalise to new examples much more effectively than if they had only seen a small subset of the data.

We have witnessed the combinatorial power of paired attributes. In real-world applications, however, we often require more than just paired attributes; we need a sequence of attributes. In natural language processing, it is necessary to construct models similar to natural language models that comprehend entire sentences. Here, the complexity dramatically escalates. For example, let's assume an English sentence comprises 20 words with a vocabulary of 40,000 words. Consequently, the number of possible sentences reaches as high as $40,000^{20} = 10^{300}$. We cannot directly manage such immense complexity.

Recently, we have seen remarkably effective methods emerge to manage this computational complexity. These include using word embeddings to transform discrete letters and words into continuous vector representations and employing Large Language Models (LLMs) for regression approximation to process new words or sentences. More effectively, all language problems are abstracted and distilled into a next-word-prediction problem, essentially a data compression challenge. Given that human language in the text can serve as training material and a benchmark, this allows for self-supervised learning, with all text on the Internet being potential training data. This approach underpins the functionality of ChatGPT and exemplifies, once again, the unreasonable effectiveness of Big Data.

While many challenges remain in the field of machine learning, the unreasonable effectiveness of data-driven methods is undeniable. We can anticipate increasingly impressive outcomes across a broad spectrum of applications in the future.

Chapter 3: AI Storytelling – A Human-Centered Perspective from Moral Philosophy

Dr. Anders Hedman
Department of Media Technology and Interaction Design
KTH, Royal Institute of Technology, Sweden

TO WATCH THIS CHAPTER VIA WEBINARS, CLICK HERE

In this chapter, we will consider storytelling from a human-centred view by exploring three forms: human, machine and hybrid storytelling and how they relate. We begin with exploring human storytelling and machine storytelling. Then, we will consider hybrid storytelling, a combination of human and machine storytelling. Finally, we will look at authentic storytelling and what it might be in AI. As we explore storytelling, it is essential to make some initial distinctions. Strictly speaking, there is only one kind of storytelling: human storytelling. We humans tell stories as part of human psychology. It is a mental activity that requires someone to do it. AIs don't have any mental reality whatsoever and so cannot tell stories. However, we can treat AIs as if they were telling stories, which we will do in this chapter. We will focus on generative AIs (GAIs) in storytelling because that is the most common form of AI storytelling. But let us consider storytelling further to see why GAIs don't strictly tell or understand stories.

3.1 INTENTIONALITY IN HUMANS AND MACHINES

To understand GAI storytelling, let us consider how we tell stories. Think about this from your perspective: when you tell stories, you use words and sentences. Furthermore, you typically understand what those words and sentences mean, and so does, hopefully, whoever reads or listens to your stories. Suppose now that we ask a GAI (Generative AI) to tell a funny story:

ChatGPT: There once was a man who bought a new boomerang. He loved his old one but decided it was time for an upgrade. So, he threw the new boomerang with great excitement. However, no matter how hard he tried, he couldn't make the old boomerang leave his hand. He had accidentally glued his hand to it while fixing a crack earlier. The real twist? He had to catch the new boomerang with his other hand while trying to unstick the old one. Quite a sight for his neighbours!

ChatGPT does not understand this story or how terrible the joke is. Why can we understand and mean things with words and sentences but not ChatGPT? Humans have intentionality (Searle, 1980), a feature of our minds that enables us to assign meaning to words and sentences. But ChatGPT doesn't have a mind or any intentionality. When we read the story above, we assign meaning to it, and the intentionality of the story is entirely relative to us.

This is an important distinction because many AI researchers claim that they are developing digital minds of code. See, for example, the well-known so-called Pause letter (Future of Life Institute, 2023) by famous AI researchers Stuart Russell, Max Tegmark, and Nick Bostrom. If it were true that AI researchers are creating digital minds as software, then perhaps we would say that ChatGPT has a mind with intentionality. I have claimed that GAIs don't have intentionality. But why couldn't GAIs have intentionality? Let us turn to Searle again. In 1980, he designed a now-famous thought experiment to demonstrate that no digital computer could have a mind by running a computer program (Searle, 1980). GAIs like ChatGPT are programs running on digital computers, so if Searle is correct, then his thought experiment applies to any GAI. Searle's thought experiment is not about the complexity or speed of digital computers and their programs; it is independent of how advanced the technology is. We can envision future technology that works on radically different principles from a digital computer, but that is another story that we will not consider further here, no pun intended.

3.1.1 The Chinese room thought experiment

Searle came up with this thought experiment in response to work on AIs that supposedly could understand stories at Yale's AI lab by Roger Schank in the later 1970s and early 1980s. This worked because you fed a simple story to the computer, and then you could ask questions about it. The computer would then respond as if it had understood (in some cases). The AI researchers at Yale thought that this demonstrated that their story "understanding" programs understood. To see why they might have thought so, we need to revisit some earlier work by Alan Turing.

Turing proposed a test in 1950 for when we could say that a machine thinks; that is now called the Turing test. In that test, an interrogator chats with a computer and a human to determine which is which, and if the interrogator is fooled repeatedly, then at some point, according to Turing, the computer can be said to think. The story of understanding computers at Yale was thought to be based on similar principles. The evidence of understanding was that they behaved as if they understood, but is such behaviour enough? Searle thought that there was a difference between someone or something behaving as if it understood and actually understanding.

Searle said to himself let the stories be in Chinese, and I will be the computer running the story understanding program; I would not understand anything. He imagines he is in a room with two slots on a wall. Paper slips with questions in Chinese come in through one slot as “input”. When a slip comes in, he looks at it and consults a rule book, an “algorithm”, to find an answer. The algorithm says how to combine Chinese symbols into output and then writes them down on a slip that he returns through the other slot as output. Suppose Searle gets so good at this that his answers cannot be distinguished from those of a native Chinese speaker. Would that mean he understands the questions in Chinese and the answers he produces? For Searle, the answer is obvious. He doesn’t understand any of that. We can imagine that he doesn’t even know the symbols are in Chinese. But a computer following the same algorithm wouldn’t understand if Searle didn't understand. So, the upshot is that no digital computer can understand by running a computer program. It doesn’t matter how fast or powerful the digital computer is or what software it is running. It may behave as if it understood like Searle in the Chinese room, but it wouldn’t actually understand anything.

3.1.2 Syntax and Semantics

What is happening in the Chinese room? How come a computer running a program cannot understand? The problem is that the computer doesn’t have semantics or meaning. The program is entirely syntactical or formal, and you cannot get to the level of semantics or meaning from syntax. That is why Searle doesn’t understand anything in the Chinese room, and Schank’s story “understanding” programs or ChatGPT could never understand anything. One could even say that the computer doesn’t have syntax as Searle would later do. We assign a syntactical interpretation, which is the software. The computer itself cannot do that because it doesn’t have intentionality. What is there to a computer without observer relative intentionality? Atoms, particles in fields of force or whatever the latest physics tells us that the ultimate constituents of reality are—nothing there has been designed to have a mind or intentionality.

3.2 Human Storytelling

Now, we can see why human storytelling differs from computer storytelling. The computer doesn’t have any understanding. So, any meaning would have to be ascribed by humans; it would have to be observer-relative and not intrinsic to ChatGPT or any other program running on a digital computer. We tell stories as part of our mental life. Human storytelling is typically also based on gut feelings, emotions, hunches, instincts, and intuitions before thinking. Our storytelling is also anchored in one way or another to our life experiences and our background. Each individual who tells stories is unique, and how stories are told in different cultures varies.

Not only does storytelling vary between cultures, but it has done so over time. The kinds of stories we tell today are often very different from those of ancient times. What we believe and value in and across cultures and over time will influence how and what we write about.

3.3 Machine Storytelling

We have all seen how GAIs can generate stories. GAIs behave as if they were intelligent and understood. We can imagine that GAIs might even pass the Turing test. We might wonder how it's possible that GAIs can produce such good stories. But we also see that they "hallucinate" sometimes, i.e., they "make things up". So, GAIs can construct mad pieces of text that are incorrect. If a human did the same, we might say the human lied, confabulated, or had gone insane.

3.3.1 Fixing hallucinations

We may wonder here why we don't just fix the hallucinations. Couldn't we implement an algorithm to keep these generative AIs from hallucinating? But it turns out it's not that easy. It seems like it's entirely unclear how we could do it. Part of the reason here, or maybe all of the reason, is that GAIs, as we have seen, don't understand what they're writing. They appear to understand what they write; they behave as if they know but don't understand. We saw this with the Chinese room thought experiment. Algorithms are not the sorts of things that could have any understanding.

3.3.2 Common sense

The challenge posed by hallucinations is getting digital computers to behave as if they have common sense, a classic AI problem. In the 1970s, AI researchers began to realise how difficult it was. The crux is that common sense is based mainly on intuitive understanding and background capacities that lie before thought and are resistant to formalisation. The researcher Douglas Lenat attempted to build a giant database with millions of facts that would give computers common sense. Still, that approach failed, and the field entered what became known as the AI winter between the 1970s and early 1990s. The field emerged from the AI winter as researchers made rapid progress with neural networks, and that trend has continued until now. However, the problem of common sense has still not been solved.

3.3.3 Wisdom

Another problem related to that common sense is that of wisdom. Since we cannot get digital computers to behave like they have common sense, we have little reason to believe that we could bring them to act as if they were wise. Furthermore, the generative AI models are based on data from virtually all of us, so any generated output is roughly an average. But it is unclear how wisdom could be seen as such an average. We tend to admire wise people because they know how to think differently and not like everybody else when needed.

3.4 Hybrid Storytelling

Let's move on to our third kind of storytelling, hybrid storytelling. Here, storytelling is an interplay between machines and humans, and it's a new kind of dynamic.

3.4.1 Machine-dominant contribution

There's the danger that the machine produces a hallucinatory story without human understanding. That's problematic if the text is used as if it were sound. It is also unfortunate for the human collaborator because it is difficult to see how she could develop intellectually. Indeed, we seem to have a regression. But it is also possible that she understands everything the GAI has produced and can filter out any hallucinations. She might also contribute to small but critical parts of the story. This would then be a situation in which it was productive for the human to work and develop as an author, even though the text size of the contribution was negligible.

3.4.2 Equal contribution

Consider a co-creating scenario where humans and GAI make roughly equal contributions. Such a scenario can be seen as balanced regarding the volume of text produced. Under normal circumstances, the human is also likely to understand the text better and develop as an author. This would be so if writing the human part contributed to understanding the GAI part. But we can imagine that the human and GAI texts are divided into distinct parts that don't have much to do with each other, and in such a situation, it may well be that the human doesn't understand the GAI parts. So, the possible problems indicated in the GAI-dominant scenario largely remain to the extent that the GAI-generated parts are not understood.

3.4.3 Human-dominant contribution

Here, we imagine that humans produce most of the story and understand what the GAI produced. Humans can see when AI hallucinates and just takes it out. It could be good for humans to learn and develop under such conditions. Still, a small GAI-produced part of the story could be crucial to understanding and grasping the text as a whole. We begin to see here that the relative amount of human-produced text is not a very good indicator of whether or not humans are using generative AI for storytelling in a good way. It can, at most, be a very rough indicator, but it is not dependable. So, we must try to understand what good storytelling might be like in other ways. This forces us to ask what actual storytelling is, authentic storytelling.

3.5 Authentic Storytelling

Authentic storytelling is, in some sense, “real” for both the audience and the author and allows for human intellectual development. Authentic storytelling requires understanding and taking intellectual ownership of the story.

3.5.1 Ownership

To be authentic, the story must be owned by the human author intellectually. The author must be able to explain the story and answer questions, demonstrating that they know what it is about. They must understand the story, how it may be further developed, and what other choices could have been made to create it. Furthermore, the author must have done sufficient work on the story to be attributed to her. The author must be able to say rightfully that she created or co-created the story.

3.5.2 Human storytelling and authenticity

As humans, we can understand and take ownership of texts or stories we produce to engage in authentic storytelling.

3.5.3 Machine storytelling and authenticity

Since the machine doesn't understand anything, it can't take ownership, and there's no possibility of authentic storytelling. All the meaning and all the understanding is in the human who reads the generated story.

3.5.4 Hybrid storytelling and authenticity

With humans and machines working together, the machine could take over, and the human may not understand the story, along with possible hallucinations. Then, we have inauthentic hybrid storytelling. If humans understand what is produced and can work with machines to filter out hallucinations and falsities and take ownership of developing as writers, then it could be authentic storytelling.

3.6 Conclusion

We have ended our overview of storytelling with generative AI technologies from a human-centred perspective. To work well from a human perspective, GAIs ought to be used to tell authentic stories. Such stories depend on genuine understanding and taking ownership.

3.7 Study Questions

1. How is human storytelling different from machine storytelling?
2. What is hybrid storytelling, and how does it differ from its human and machine counterparts?
3. What are some ethical challenges of using generative AI in storytelling?
4. What is authentic storytelling? Why is it important?
5. How can storytellers maintain authenticity in their narratives when incorporating AI technologies?
6. What are the practical implications of hybrid storytelling? What are some benefits and drawbacks?
7. What might be some potential future advancements in AI-supported storytelling? How do you think the field might develop in the future?

Chapter 4: The Visualisation of Big Data

Ms Monika Zdanavičiūtė
Prof Tomas Krilavičius
Faculty of Informatics
Vytautas Magnus University, Kaunas, Lithuania

TO WATCH THIS CHAPTER VIA WEBINARS, [CLICK HERE]

Effective data visualisation[1] is a cornerstone of data analysis, providing a powerful tool for uncovering patterns, trends, and insights hidden in complex data sets. There exists a wide variety of different visualisation techniques that can be classified in various ways. The diversity of approaches is illustrated in Figure 4.1. An interactive version of this image can be viewed on this website.

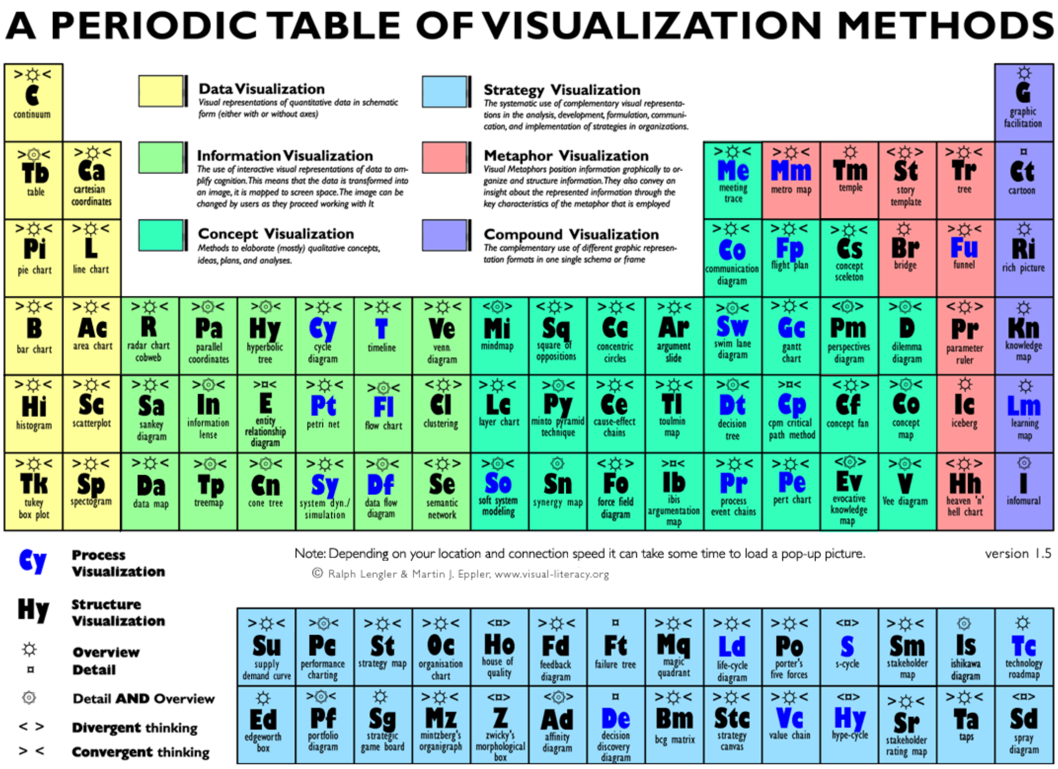

Figure 4.1: Periodic table of visualisation methods (Source: Lengler and Eppler [2007]).

This chapter presents visualisation techniques adapted to reveal different aspects of information. The beginning of the chapter delves into the basic techniques that provide a solid

[1] **Data visualisation:** Data visualisation is the process of turning numbers and data into pictures and graphs to make it easier to understand and see patterns. On the other hand, 'effective data visualisation' involves designing charts, graphs, and other visual tools clearly and insightfully, making complex data easily understandable at a glance, highlighting key trends, patterns, and outliers without overwhelming the viewer.

foundation for visual representation. Next, multidimensional data[2] representation and direct methods are analysed to reveal relationships between multiple variables[3]. In addition, we will explore the transformative effect of projections, a technique that allows us to transform complex multidimensional data into forms that are easier to understand and reveal meaning. Together, these techniques form a comprehensive toolkit that empowers analysts and researchers to transform raw data into compelling visual narratives that facilitate more profound understanding and informed decision-making.

4.1. Basic visualisation methods

In this section, we will explore various basic visualisation methods, each designed to effectively communicate data and insights through graphical representation.

4.1.1. Basic (Line) charts

A line chart shows data points connected by straight line segments. It is commonly used to show trends and patterns in data over a continuous interval/period. In a line graph, the horizontal axis usually represents the independent variable[4] (such as time or categories), and the vertical axis represents the dependent variable[5] (such as the measured data). A marker represents each data point on the chart, which are connected by lines to illustrate the overall trend or pattern in the data. Line charts effectively show how a variable changes over time, identify patterns and compare multiple datasets.

Line charts are handy in scenarios where the data points are ordered[6], and their relationships are essential. They are commonly used in various fields, including finance, economics, science, and engineering, to visualise and analyse trends in data.

[2] **Multidimensional data:** Imagine you have a treasure chest with different layers: on one layer, you have coins; on another, you have gems; and on another, you have maps. Multidimensional data is like this chest: it has information (treasure) sorted into different layers (dimensions) so you can find exactly what you need, whether it's coins, gems, or maps, and see how they all relate to each other.

[3] **Variable:** A variable is anything that can change or vary, like temperature, age, or income, in an experiment or data collection.

[4] **Independent variable:** An independent variable is a factor you adjust in an experiment to see how it influences the outcome, such as changing the amount of sunlight plants receive to see how it affects their growth.

[5] **Dependent variable:** Based on the example given, a dependent variable is the outcome observed and measured in an experiment, such as the growth of plants, which changes in response to the amount of sunlight they receive.

[6] **Ordered data:** Ordered data is a type of data that is arranged in a specific, meaningful sequence, either ascending or descending, based on numerical or categorical values.

Only dots have a direct meaning if nominal values are depicted, while lines (straight lines, splines, regression) depict changes. In Figures 4.2 and 4.3, you can see examples of a line graph (straight and splines) illustrating the changes over time in the number of scientists per 1,000,000 people.

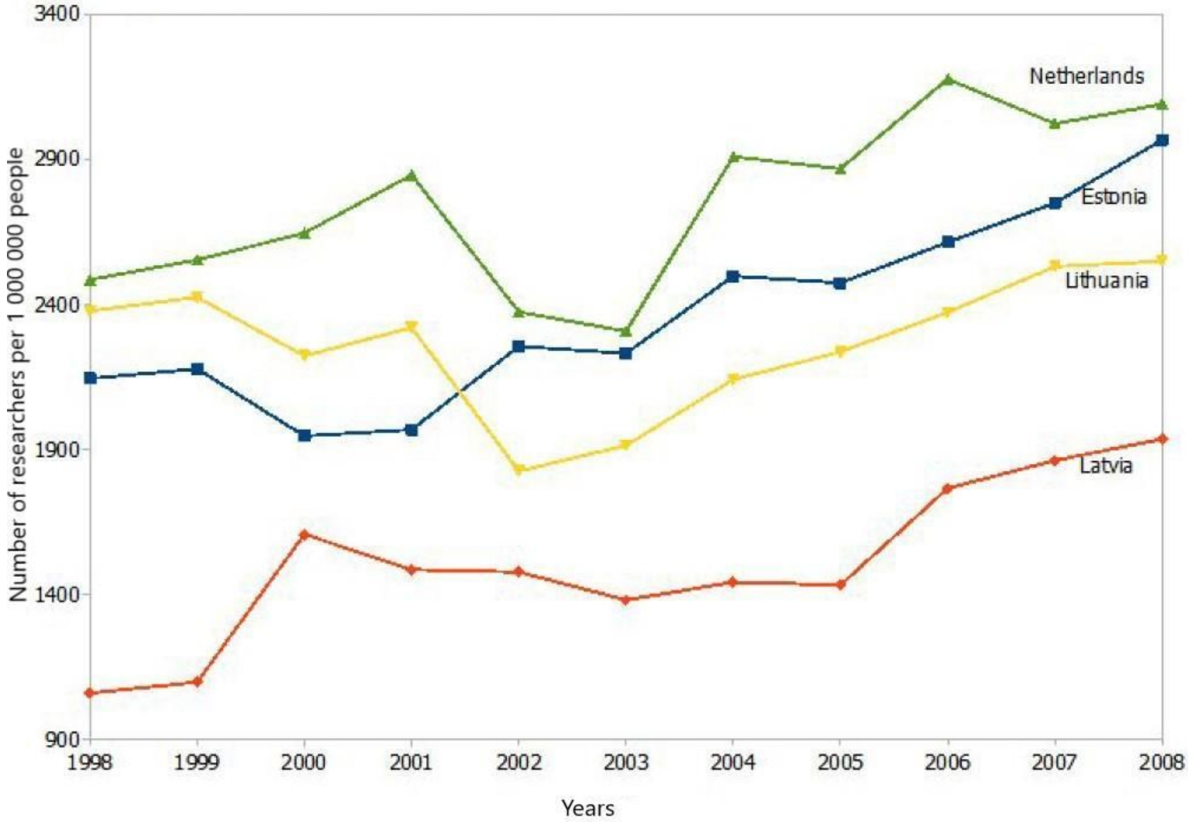

Figure 4.2: Number of scientists per 1 million, straight lines.

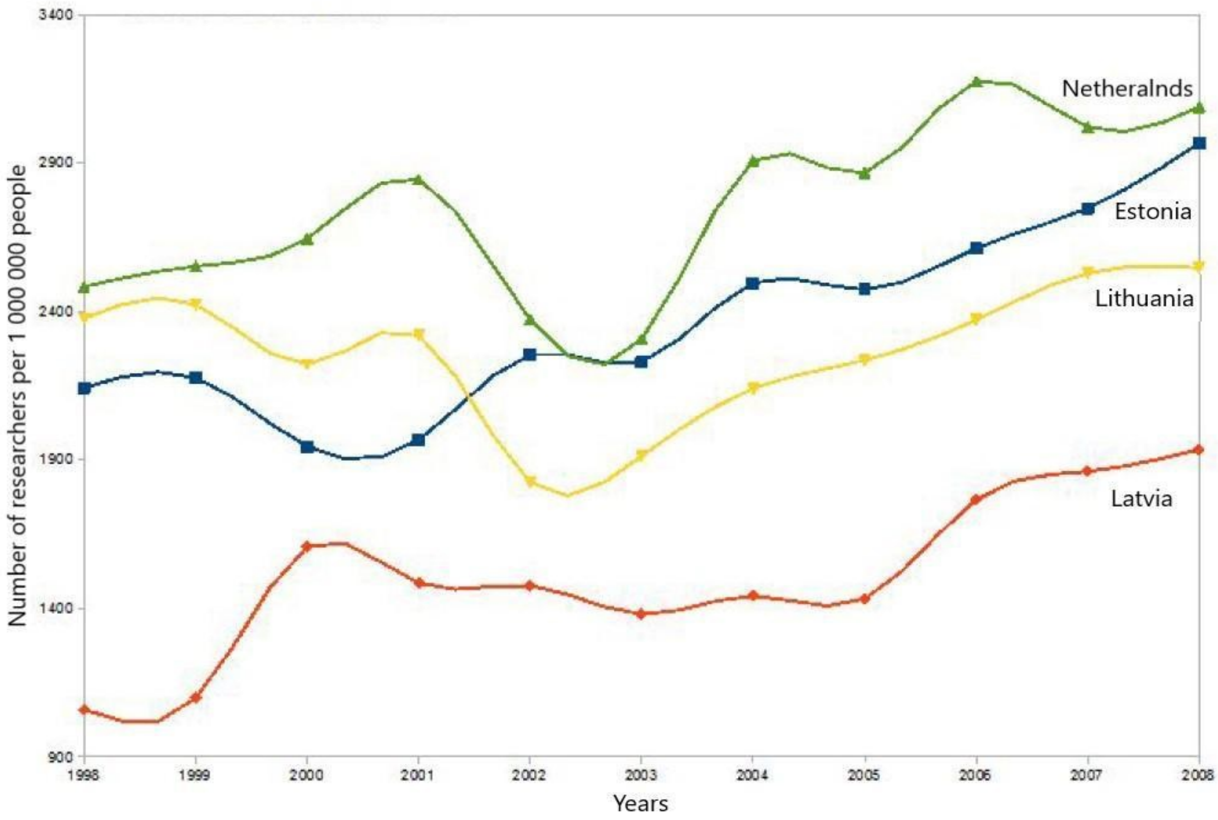

Figure 4.3: Number of scientists per 1 million, splines.

4.1.2. Area Charts

An area chart is a type of data visualisation similar to a line chart, but the area below the line is filled with colour or shading. Like a line chart, it is beneficial for showing trends over a continuous interval or period. The area between the line and the horizontal axis (x-axis) is usually filled with colour or pattern to emphasise the magnitude of values over time or across categories.

Area charts are effective for illustrating cumulative values[7] and show the general pattern or distribution of data. They are most often used when emphasising the overall size of a set of values is essential, such as tracking the total income of several months or comparing the total contribution of different categories to the total amount.

[7] **Cumulative values:** Cumulative values are like a running total, where you keep adding each new number to the sum of all the previous numbers.

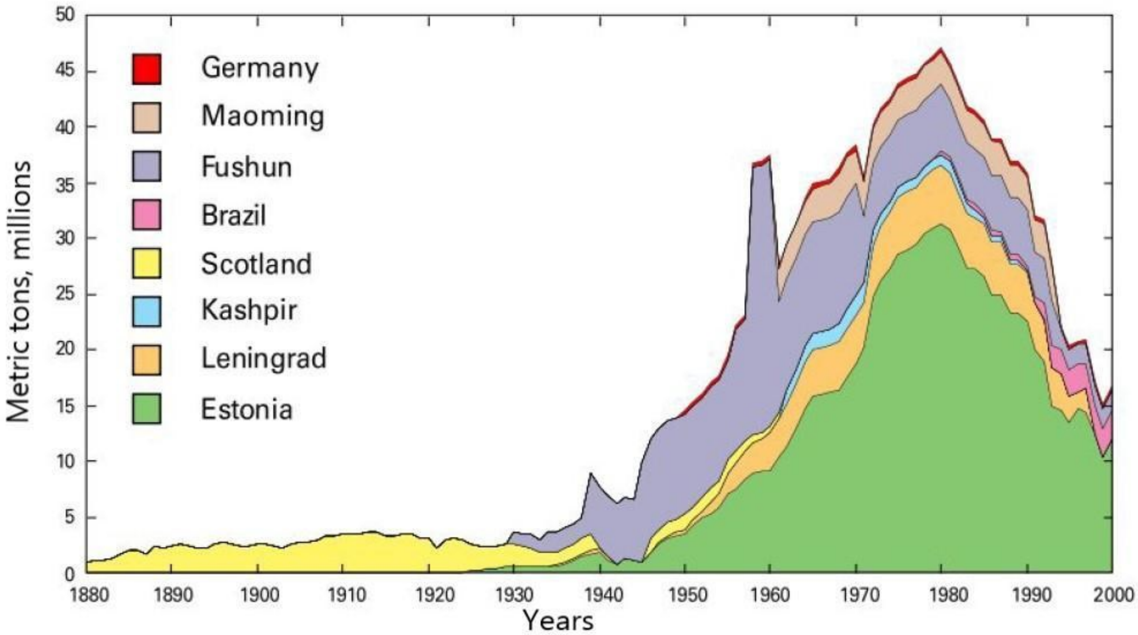

Figure 4.4: Production of oil shale, area graph.

In area charts, the x-axis (horizontal) usually represents the independent variable (such as time or categories), and the y-axis (vertically) represents the dependent variable (measured data). The shape and colour of the filled area make it easy to visualise changes and variations in the data. See Figure 4.4, which depicts oil shale production using an area graph.

4.1.3. Bar charts

A bar chart is a common type of data visualisation that presents categorical data with rectangular bars. The lengths of these bars are proportional to the values they represent. Bar charts compare and display the relative sizes of different categories or groups. The categories are typically shown on the horizontal axis (x-axis), while the values or frequencies they represent are displayed on the vertical axis (y-axis).

Bar charts are versatile and can be used for various purposes, such as comparing individual values, showing trends over time, and illustrating data distribution. They are widely used in different fields, including business, economics, science, and social sciences, to convey information visually effectively.

There are two main types of bar charts:

➢ **Vertical Bar Chart:** The bars extend vertically from the horizontal axis in this type. Each bar represents a category, and its height corresponds to the value it represents (examples can be seen in Figures 4.5 and 4.6).

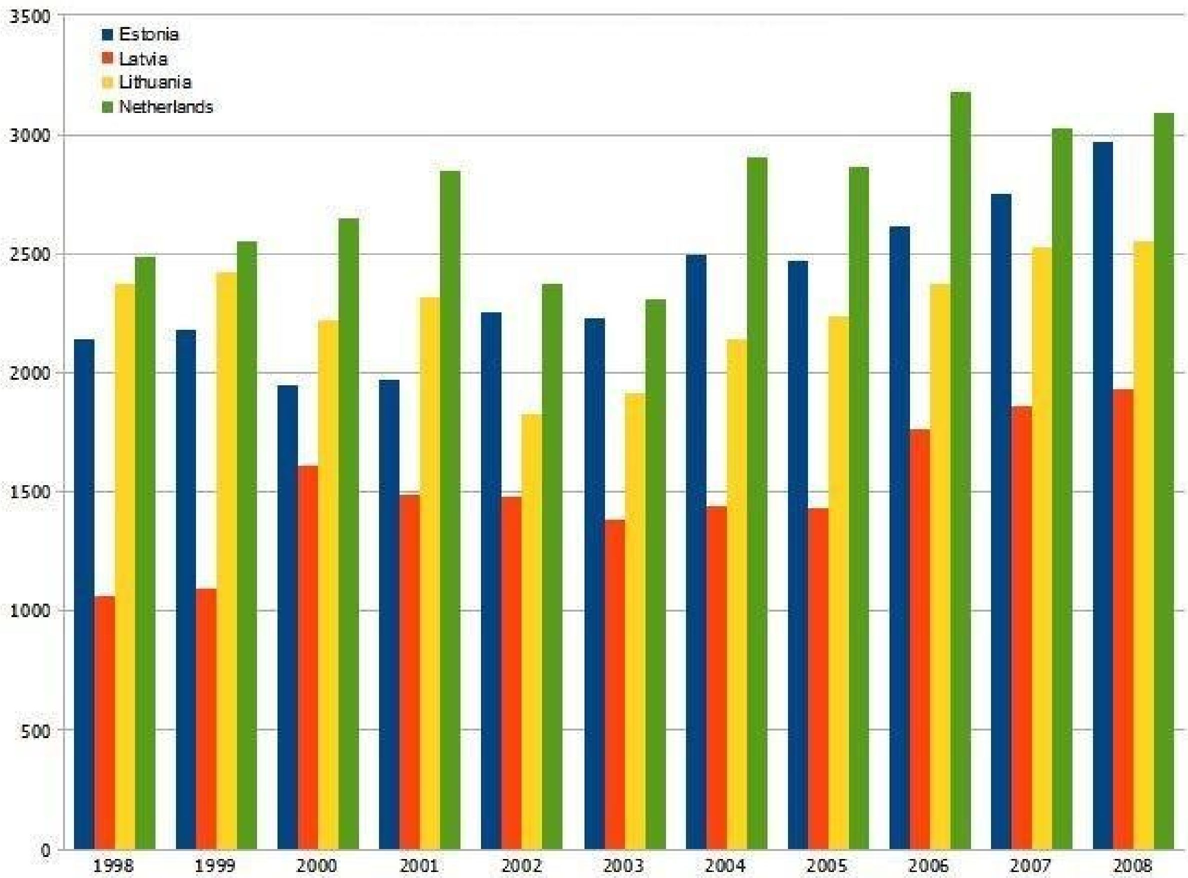

Figure 4.5: Number of scientists per 1 million, bar chart.

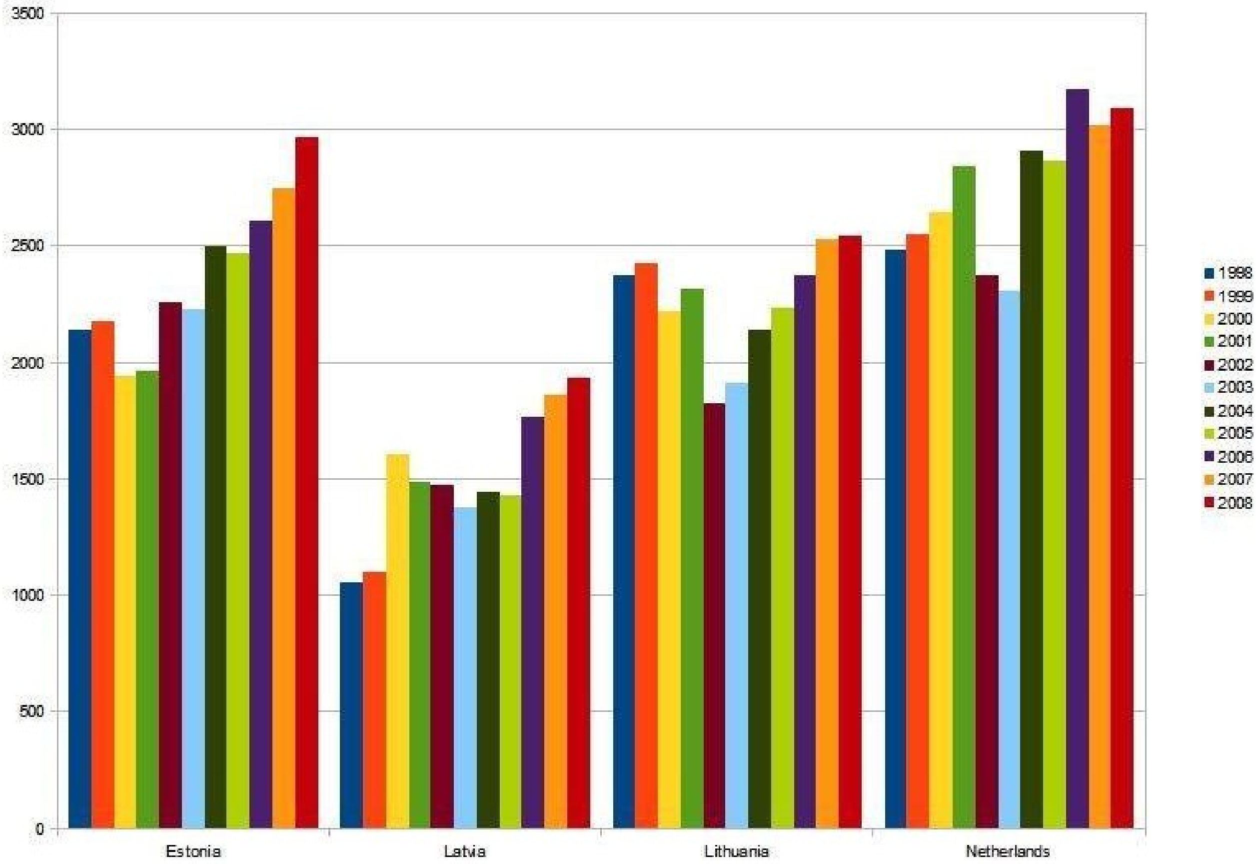

Figure 4.6: Number of scientists per 1 million, bar chart, different grouping.

- **Horizontal Bar Chart:** In this type, the bars extend horizontally from the vertical axis. Each bar represents a category, and its length corresponds to the value it represents (an example can be seen in Figure 4.7).

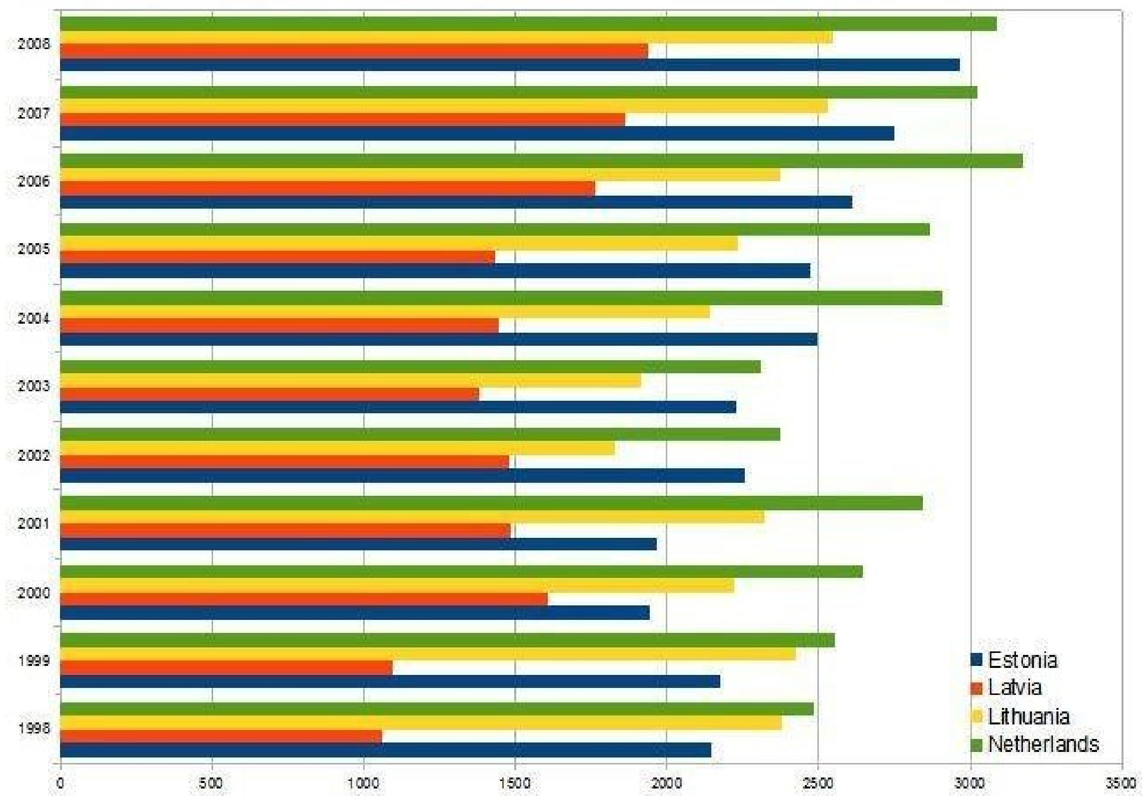

Figure 4.7: Number of scientists per 1 million, rotated bar chart.

Sometimes, 3D bar graphs are used, but usually, they only make the situation worse and do not represent the information better. Examples of 3D bar graphs are shown in Figures 4.8 and 4.9.

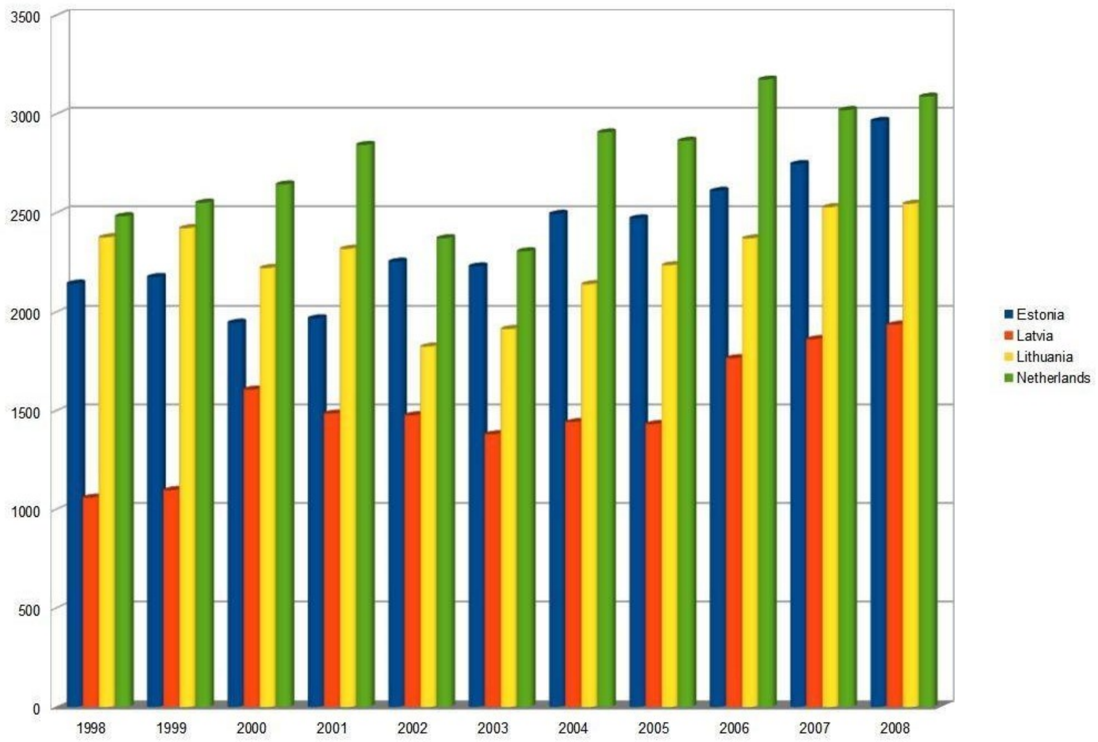

Figure 4.8: Number of scientists per 1 million, 3D bar chart.

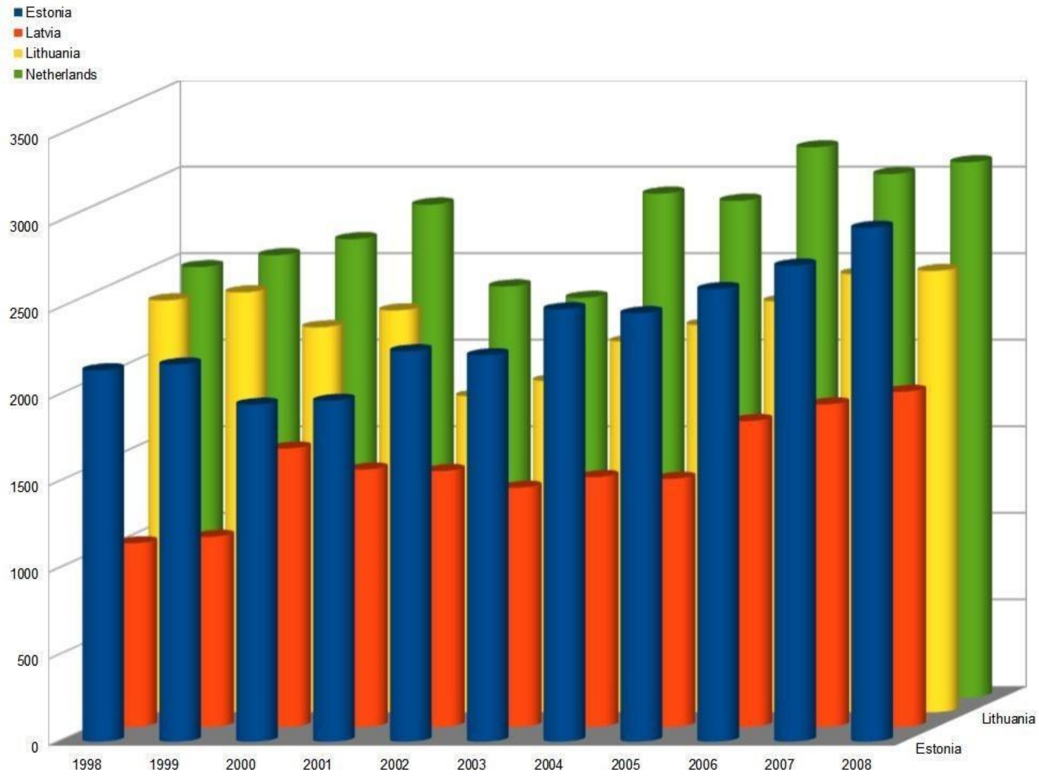

Figure 4.9: Number of scientists per 1 million, 3D bar chart in 3D.

4.1.4. Histograms

A histogram is a graphical representation of the distribution of a dataset. It visualises the underlying frequency distribution of a discrete or continuous variable[8]. In a histogram, the data is divided into intervals or bins, and the height of a bar represents the frequency (or number/count) of observations falling into each bin/interval.

Key features of a histogram include:

- **Bins:** Intervals along the horizontal axis that divide the range of the data. Each bin represents a range of values.
- **Frequency:** The number of data points falling within each bin. The height of the bars represents this.
- **Bars:** Vertical columns or rectangles above each bin, illustrating the frequency or count of data points within that bin.

Examples of histograms are depicted in Figures 4.10 and 4.11.

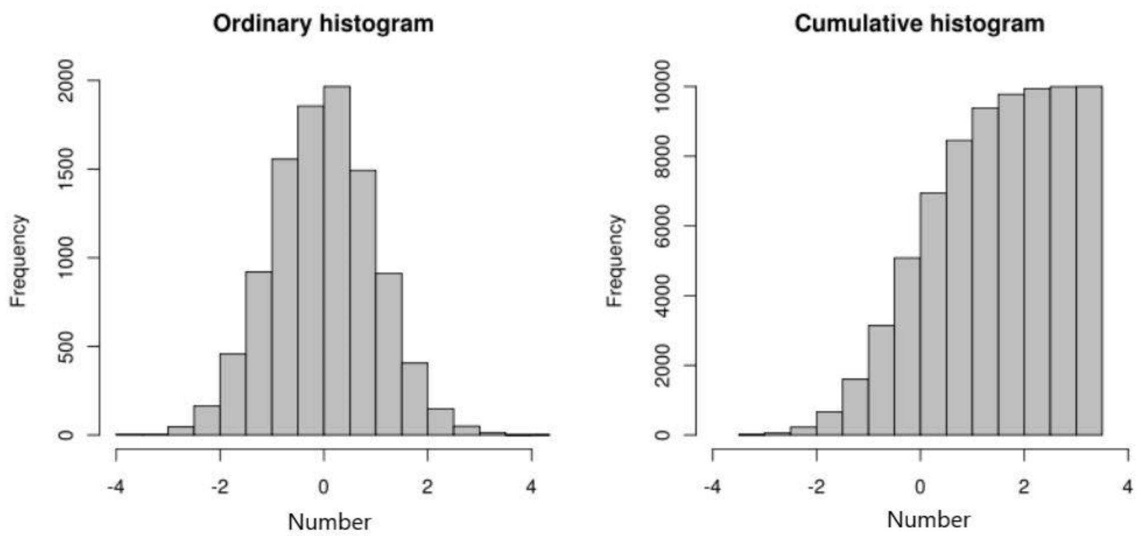

Figure 4.10: An ordinary and cumulative histogram of the same data. The data shown is a random sample of 10,000 points from a normal distribution with a mean of 0 and a standard deviation of 1 (Source: [Wikipedia](#)).

[8] **Discrete & continuous variables:** Discrete variables represent counts (e.g. the number of objects in a collection). Continuous variables represent measurable amounts (e.g. water volume or weight).

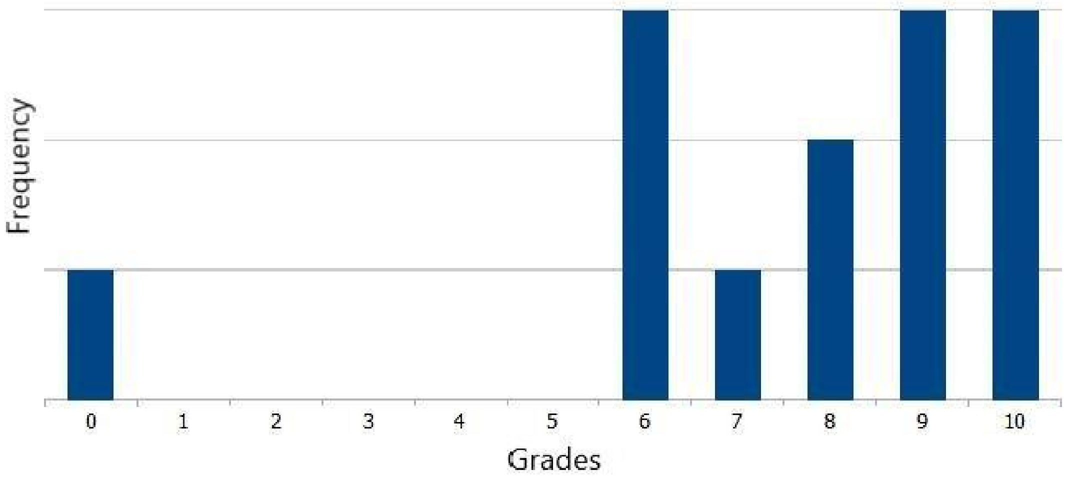

Figure 4.11: Distribution of grades, histogram.

4.1.5. PIE CHARTS

A pie chart is a circular statistical graph divided into sections to show numerical proportions. Each piece is a proportional part of the whole, and the sum of all the pieces is 100%. Pie charts are commonly used to show the distribution of categorical data[9] or the relative sizes of different components within a whole.

Key features of a pie chart:

- **Slices:** Each category or component is represented by a slice of the pie. The size of each section is proportional to the percentage of the whole it represents.
- **Centre Point:** The centre of a pie chart is usually left blank so that the lobes are visible.
- **Labels:** Categories are often labelled directly in each column or a legend behind the pie chart.

The pie charts in Figures 4.12 and 4.13 represent the same information using different pie chart variations. The former is a poor representation of the data as the numbers are not visible, and what they mean is unclear. At the same time, the latter is easier to understand due to the explanations provided.

[9] **Categorical data:** Categorical data sorts things into labels or groups, like types of music or colours, without numbers involved.

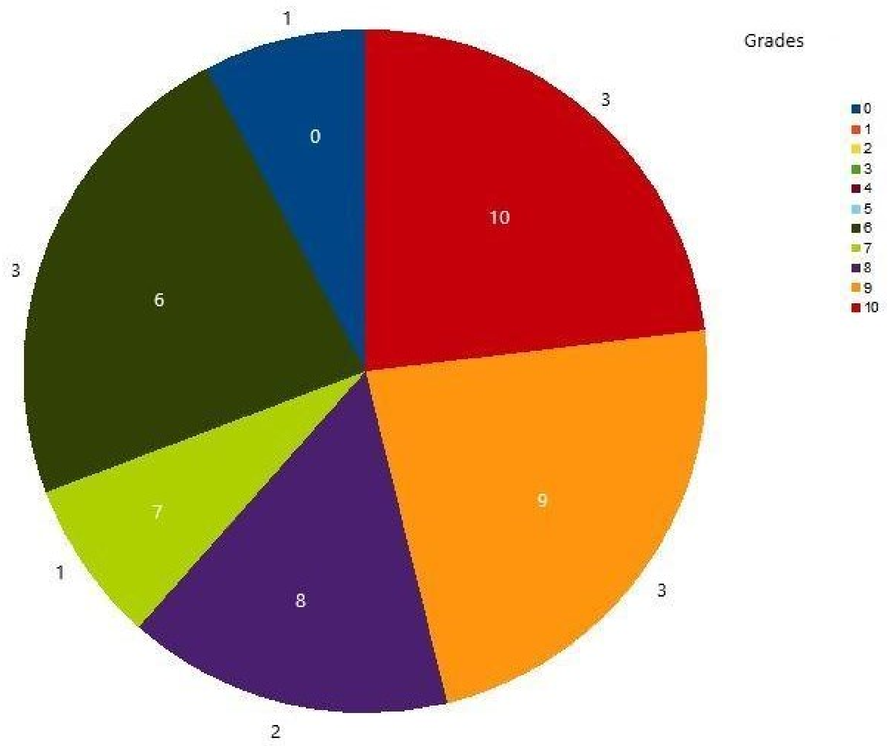

Figure 4.12: Distribution of Grades, pie chart 1.

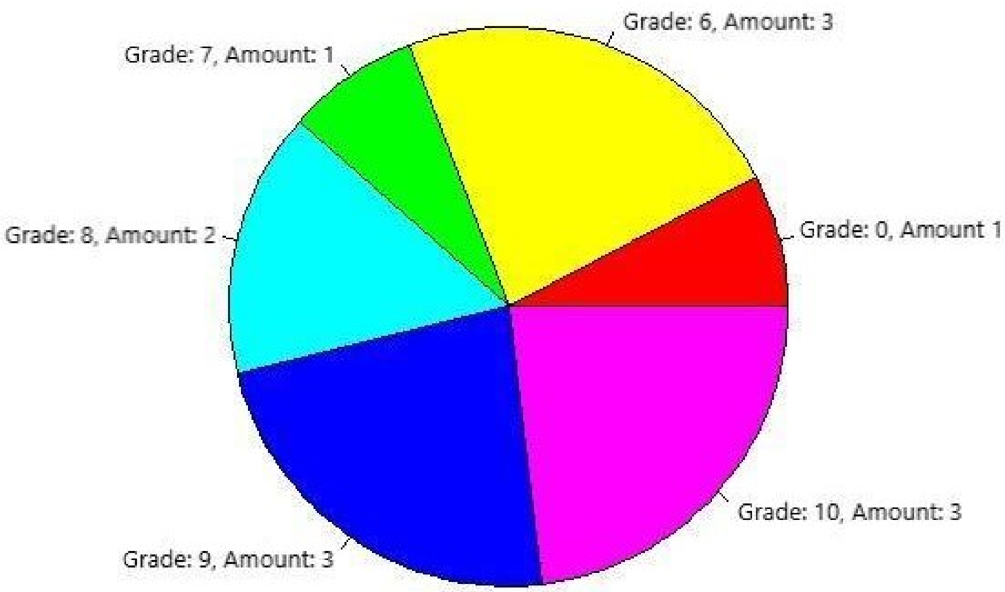

Figure 4.13: Distribution of Grades, pie chart 2.

However, pie charts may not be as effective when there are too many categories or when the differences in proportions are subtle. When choosing a visualisation method, it is imperative to use pie charts judiciously and consider the context and nature of the data.

4.1.6. Box plots

Box plots represent a class of different plots, such as box plots, violin plots, and others commonly used in statistics, which graphically represent the distribution of a dataset. It displays the summary statistics of a dataset:

1. Minimum
2. Lower quartile (quarter-point), the first quartile (Q1), 25% of data
3. Median, the second quartile (Q2), 50% of data
4. Upper quartile, the third quartile (Q3), 75% of data
5. Maximum

Key components of a box plot include (see Figure 4.14):

- **Box:** The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The height of the box indicates the spread of the middle 50% of the data.
- **Whiskers:** Lines extending from the box to the minimum and maximum values within a specified range. The length of the whiskers provides information about the spread of the entire dataset.
- **Median Line:** A line inside the box representing the dataset's median (Q2).
- **Outliers:** Individual data points outside the whiskers are displayed as individual points.

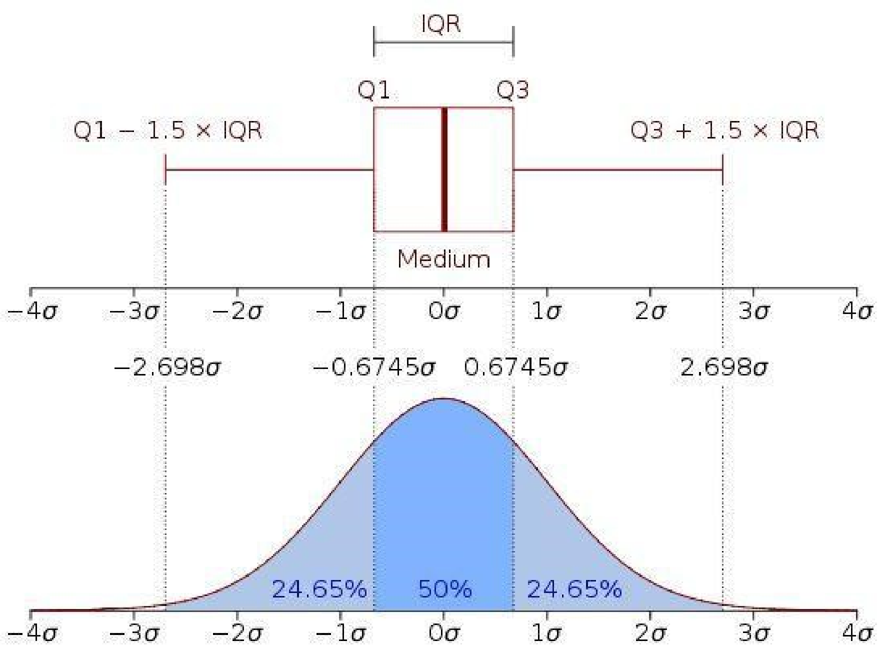

Figure 4.14: Structure of box plot (Source: Wikipedia).

Box plots are particularly useful for comparing the distribution of different datasets or identifying the presence of outliers. They provide a visual summary of the range, central tendency, and dispersion of data in a concise and easily interpretable manner. Box plots are commonly used in statistics, data analysis, and research to gain insights into a dataset's characteristics. Figures 4.15 and 4.16 depict several different box plots.

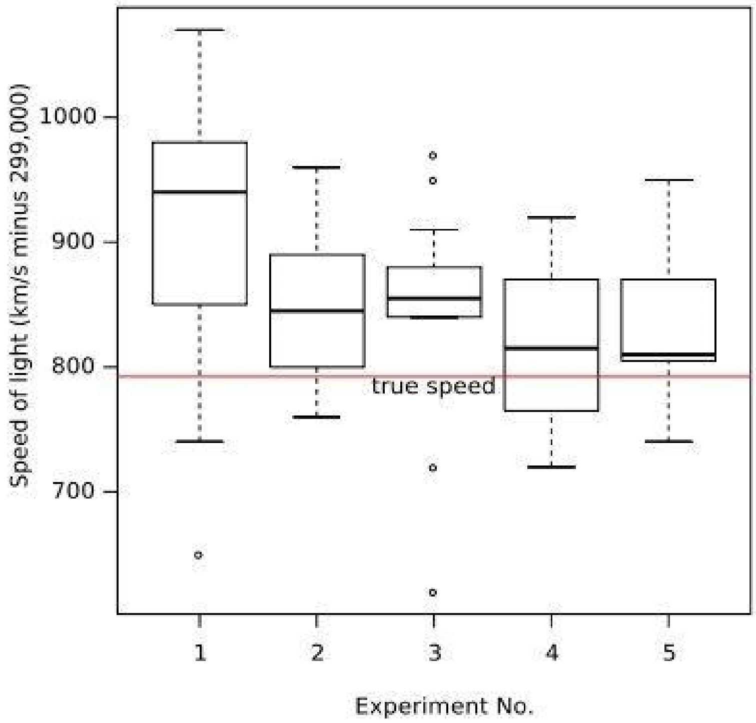

Figure 4.15: Results of the speed of light experiments, box plot (Source: Wikipedia).

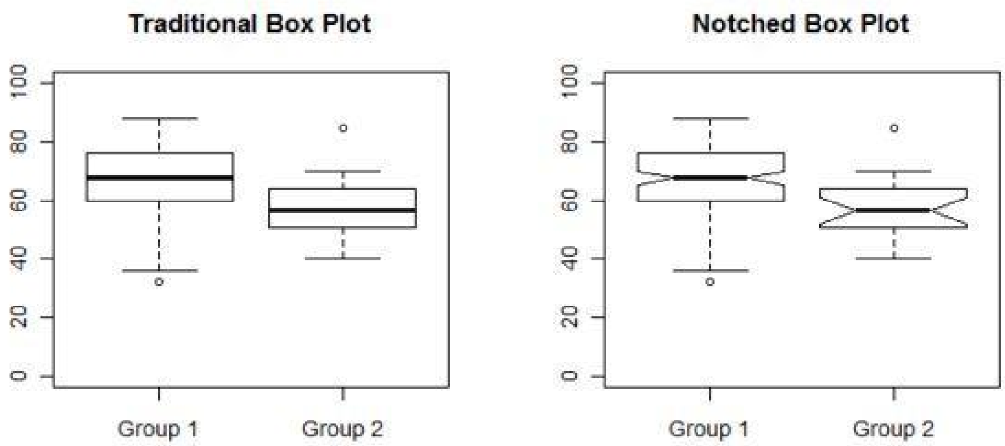

Figure 4.16: Examples of different types of box plots (Source: Wikipedia).

4.2 MULTIDIMENSIONAL DATA STORYTELLING: DIRECT METHODS

Multidimensional data refers to data that is organised and presented in more than one dimension. In simpler terms, it involves data that has multiple attributes, variables, or functions. Each dimension represents a different aspect or characteristic of the data, and combining these dimensions provides a more complete picture of the data set (Figure 4.17).

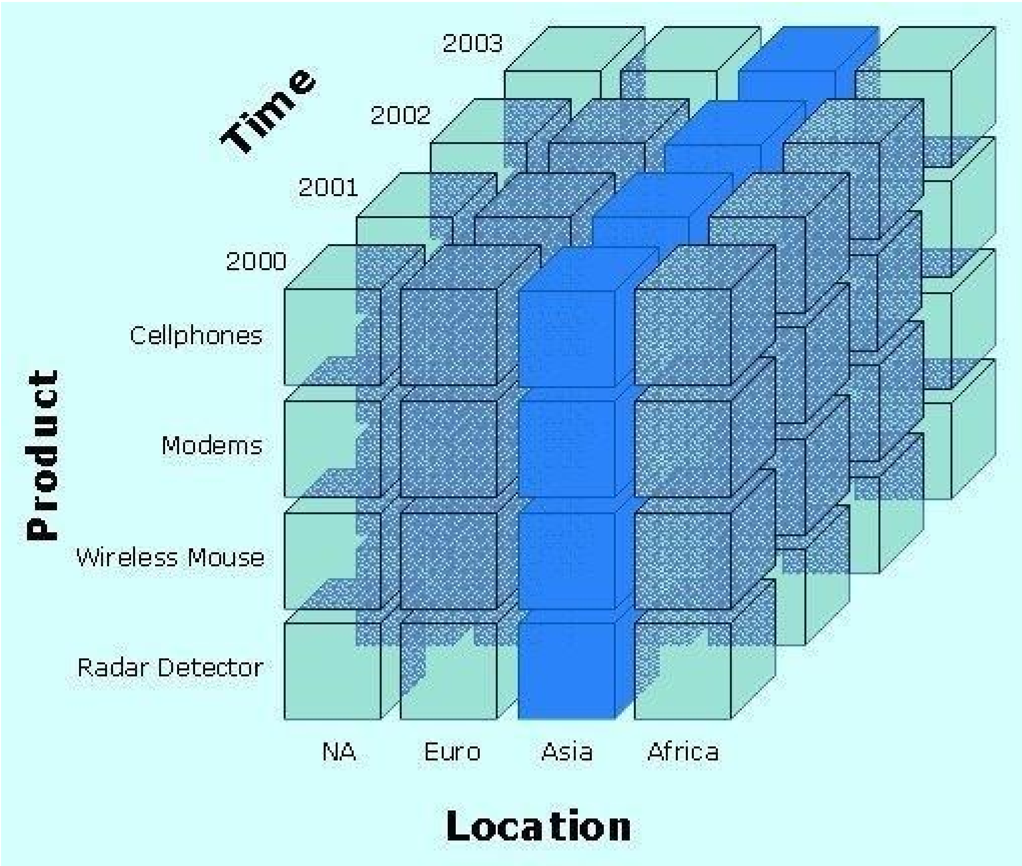

Figure 4.17: Multidimensional data visualisation.

Multidimensional data visualisation objectives are:

- **Research** – sometimes, unique graphs are proper only for the data research process, but their usage for data presentation may be unsuccessful. Usually, such graphs may be informative, present exciting information, and extract essential aspects of the data.
- **Represent** – after the data research is completed and the information from the data is removed, the visual representation of the results follows.

Various authors present different classifications of visualisation methods. One of the classifications of multidimensional data visualisation techniques (Dzemyda et al., 2008):

A. Direct visualisation methods – when all attributes of the multidimensional object are presented visually.

a) Geometric techniques

i. Scatter plots
ii. Matrix of scatter plots

iii. Line/multiline graphs
iv. Permutation matrix
v. Survey plots
vi. Andrew's curves
vii. Parallel coordinates
viii. Radial visualisation

b) Iconographic display

c) Hierarchical display

B. Dimension reduction methods – maps multidimensional data into lower dimension space.

a) Linear projection techniques

i. Principal component analysis
ii. Linear discriminant analysis
iii. Projection pursuit

b) Nonlinear projection techniques

i. Multidimensional scaling
ii. ISOMAP
iii. Locally linear embedding
iv. Principal curves
v. Triangulation
vi. Artificial neural network

Next, in this section, some direct methods for visualising multidimensional data are presented.

4.2.1 Scatter plot

A scatter plot is a data visualisation type showing individual data points in a two-dimensional graph. Each point in the graph represents the values of two variables. One variable is plotted on the x-axis (horizontal axis), and the other on the y-axis (vertical axis) (Figure 4.18). Scatter plots are useful for visually examining the relationship or correlation between two continuous variables.

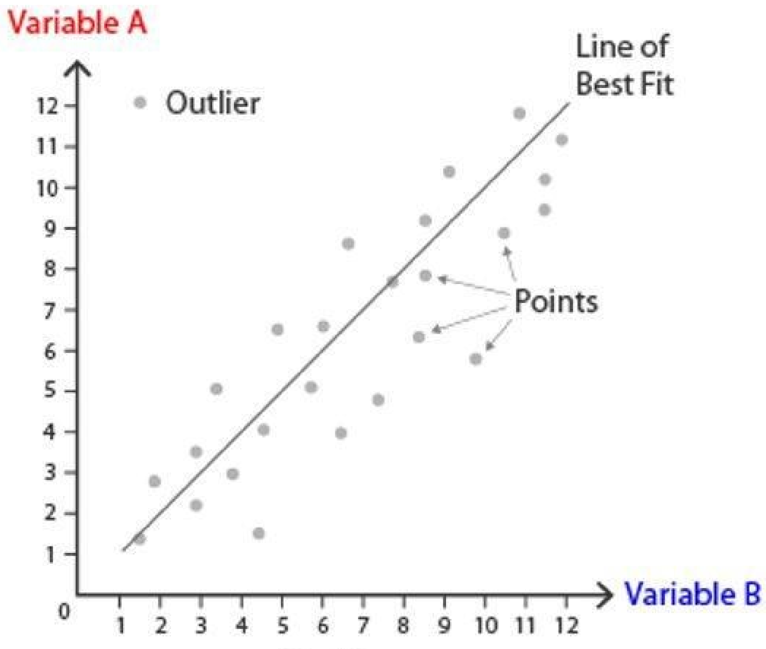

Figure 4.18: Structure of scatter plot (Source: datavizcatalogue).

Key features of a scatter plot include:

- **Data Points:** Each data point on the plot represents a specific observation or data entry with values for both the x and y variables.
- **X-Axis and Y-Axis:** The horizontal axis (x-axis) typically represents the independent variable, and the vertical axis (y-axis) represents the dependent variable.
- **Trend Line:** Sometimes, a trend line (also known as a regression line) is added to the scatter plot to illustrate the general direction or pattern of the relationship between the variables.

Scatter plots are particularly useful for identifying patterns, trends, and the strength of the relationship between two variables. The following scenarios can be observed in a scatter plot:

- Positive correlation: Points generally trend upward.
- Negative correlation: Points trend typically downward.
- No correlation: Points are scattered randomly with no apparent pattern.

Figure 4.19 shows several different trends represented by scatter plots. Figure 4.20 shows how the number of accidents with other animals is related to the number of people killed.

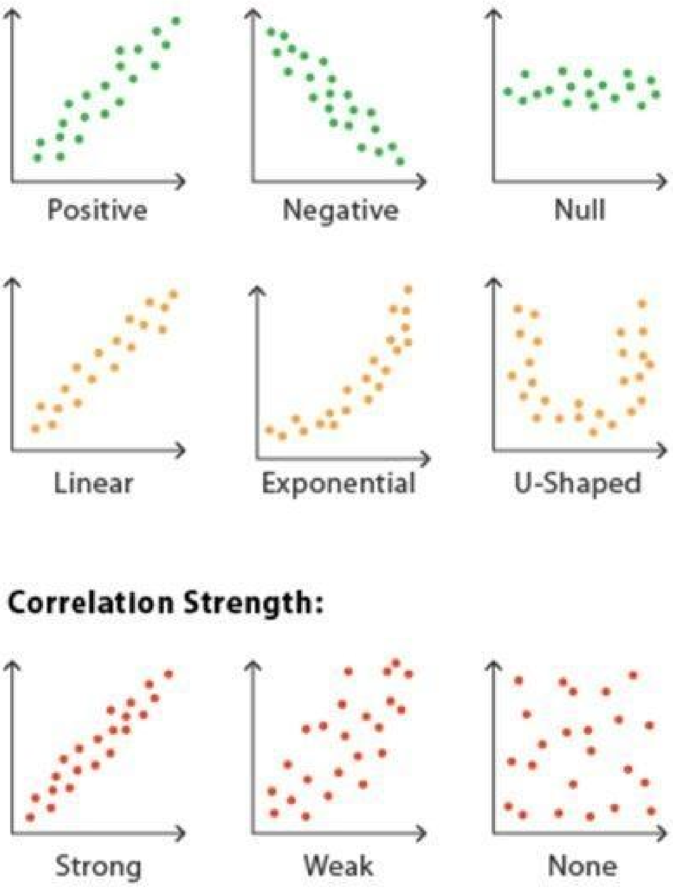

Figure 4.19: Different trends visualised by scatter plot (Source: datavizcatalogue).

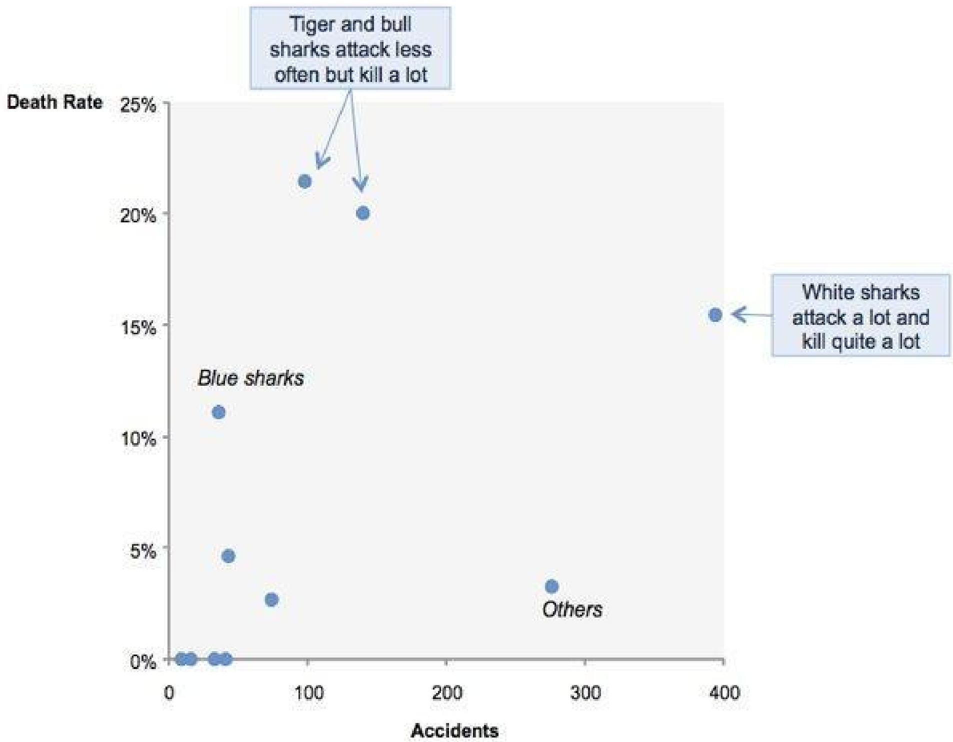

Figure 4.20: Number of accidents and death rate correlation, scatter plot.

4.2.2 Heat maps

A heatmap is a graphical representation of data in which the values in a matrix are represented by colour. It is an effective way to visualise the size of a phenomenon in two-dimensional space. Heatmaps are commonly used to describe the intensity of values, such as the concentration of data points or the correlation between variables, by assigning different colours to different levels of magnitude. Usually, heat maps are used to depict clusters. Figure 4.21 shows an example of heatmaps, whereby displaying values in colours, objects can be separated into specific clusters.

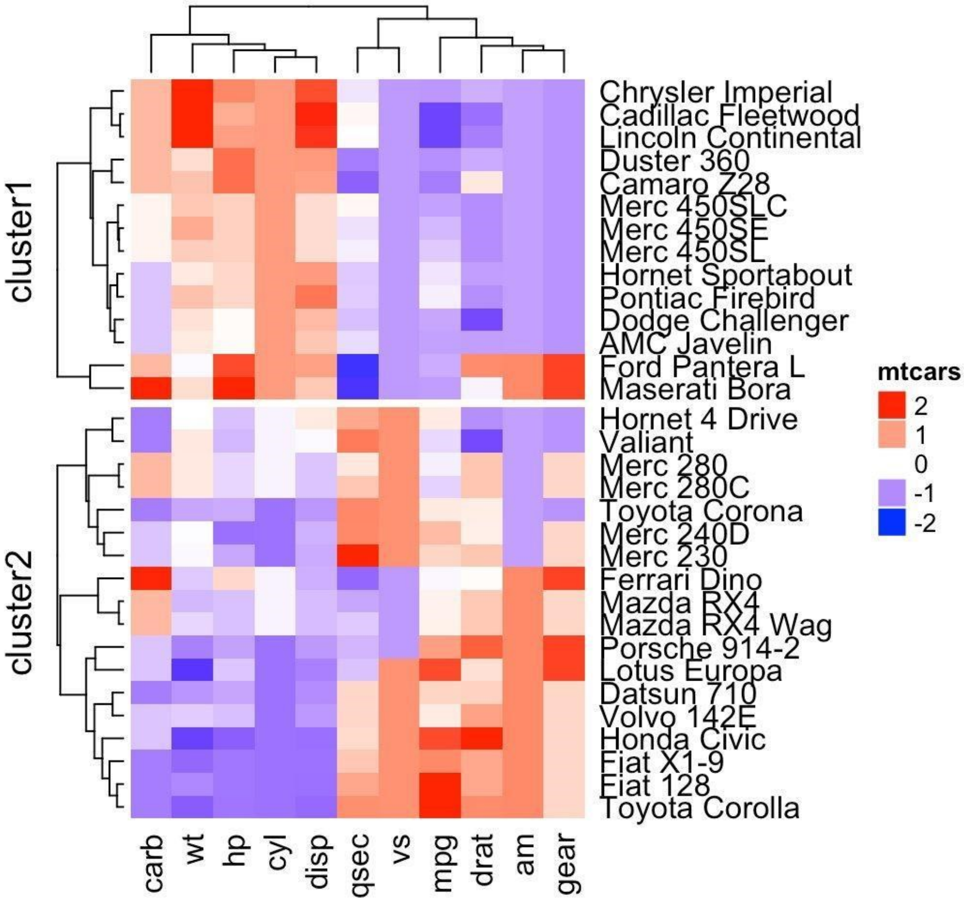

Figure 4.21: Different cars and their various technical parameter values, heat map.

Key features of a heat map include:

- **Colour Gradient:** A colour gradient is applied to the cells of a matrix, where different colours represent different values or intensity levels. Typically, a spectrum of colours from cool to warm is used, with cool colours (e.g., blue) representing lower values and warm colours (e.g., red) representing higher values.

➢ **Matrix Representation:** Data is organised in a matrix format, with rows and columns corresponding to different categories or variables. Each cell in the matrix is assigned a colour based on the value it represents.

➢ **Legend:** A legend is often included to provide a reference for interpreting the colours in the heat map.

4.2.3 GRAPHS

The graph is a data structure that consists of a set of nodes (vertices) and a set of edges connecting pairs of nodes. Graphs are widely used to model relationships between entities. Different types of graphs exist, including directed graphs (digraphs), undirected graphs, weighted graphs, and more (Figure 4.22).

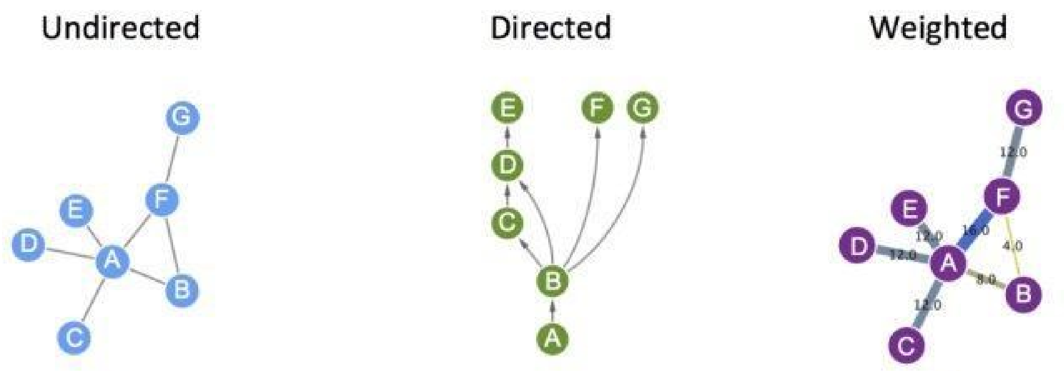

Figure 4.22: Graph types (Source: EMBL-EBI).

A specific type of graph called a "tree" is a particular case of a graph with specific properties. In a tree:

➢ **No Cycles:** The graph has no cycles or loops, meaning you cannot start at a node and follow edges to return to the same node.

➢ **Connected:** Every pair of nodes is connected by exactly one path.

➢ **Single Root Node:** There is one designated root node from which all other nodes are reachable.

➢ **Hierarchical Structure:** The nodes form a hierarchy, and each node (except the root) has exactly one parent but can have multiple children.

➢ **Acyclic:** Trees are acyclic, meaning there are no closed loops or circuits

An example of a tree is depicted in Figure 4.23.

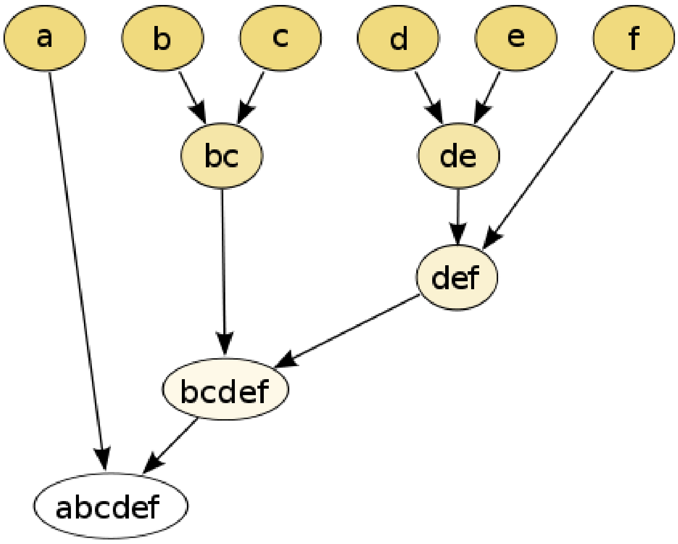

Figure 4.23: Tree (specific graph).

4.2.4 MOSAIC PLOT

A mosaic plot is a data visualisation technique that shows the relationship between two categorical variables in a two-dimensional space. It is a graphical representation of contingency tables that allows you to visualise the distribution of categories and their relationships.

In a mosaic plot:

- The width of the columns represents the distribution of one categorical variable.
- The height of the rows represents the distribution of another categorical variable.
- The area of each rectangle within the plot is proportional to the joint frequency of the corresponding combination of categories.

The plot consists of a grid of rectangles, where each rectangle represents a combination of categories from the two variables. Colours or shading within the rectangles can emphasise differences in frequencies or proportions. An example of a mosaic plot is shown in Figure 4.24.

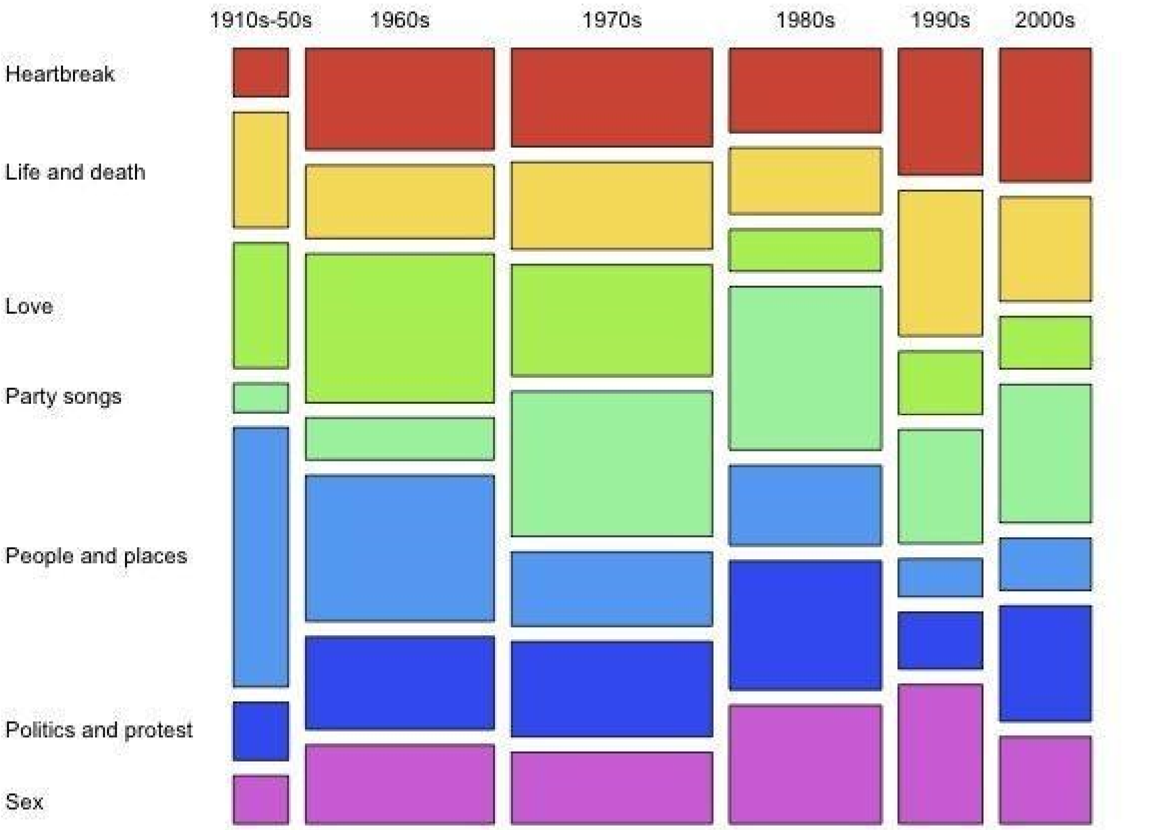

Figure 4.24: Distribution through a time of different musical themes in the Guardian's list of “1000 songs to hear before you die”, mosaic plot (Source: Wikipedia).

4.2.5 SANKEY

A Sankey diagram, also known as a Sankey diagram or flowchart, is a specific type of data visualisation that illustrates the flow of resources or information between multiple entities. It is particularly effective in showing the distribution and direction of resources such as energy, money or materials through a system or process. Sankey diagrams use directional arrows to represent flow, and the width of the arrows is proportional to the quantity represented.

Key features of a Sankey diagram include:

- **Nodes:** Represent entities or stages in a system, such as sources, processes, or destinations.
- **Arrows:** Represent the flow of resources or information between nodes. The width of the arrows corresponds to the quantity being transferred.
- **Direction:** Arrows typically flow from one node to another, indicating the direction of the flow.
- **Quantities:** The arrows' thickness or the bands' width is proportional to the quantity of the flow.

An example of the Sankey diagram is presented in Figure 4.25 (an interactive version of this diagram can be found here)

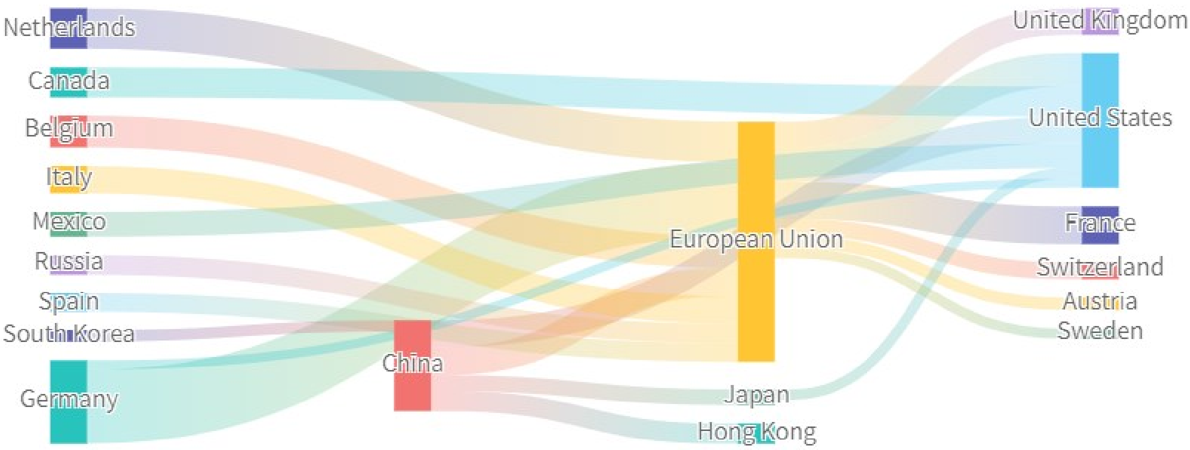

Figure 4.25: Immigrant flows visualisation using Sankey diagram (Source: Fusion charts).

4.2.6 Parallel coordinates

A parallel coordinate plot is a data visualisation technique representing multidimensional data in a two-dimensional space. It is particularly effective for studying relationships and patterns in data sets with multiple variables. In a parallel coordinate plot, each variable is represented by a vertical axis (separate vertical axis for each feature), and individual data points are connected by lines that run parallel to these axes. It is important to note that this type of graphing can only be helpful for data with a limited number of attributes and observations.

Key features of a parallel coordinate plot include:

- **Vertical Axes:** Each variable is assigned a vertical axis. These axes run parallel to each other.
- **Lines:** Data points are represented as lines that connect to each axis. The position of a line on each axis corresponds to the variable's value.
- **Interaction:** Users can visually follow the lines across multiple axes to observe patterns, trends, or relationships between variables.
- **Brushing and Linking:** This interactive feature allows users to select a range on one axis, highlighting the corresponding lines on other axes. This makes it easier to identify patterns within a subset of the data.

A parallel coordinate plot example is presented in Figure 4.26.

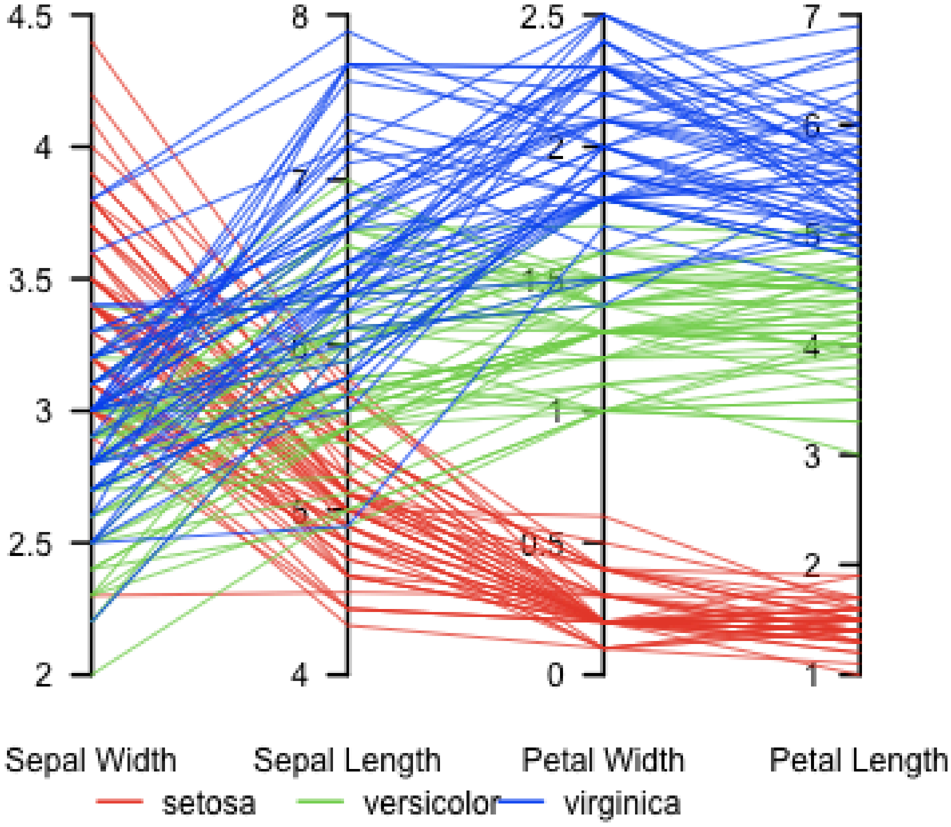

Figure 4.26: Fisher's Iris data, parallel coordinates plot (Source: Wikipedia).

4.3 MULTIDIMENSIONAL DATA STORYTELLING: PROJECTIONS

This section is aimed towards students with a more established knowledge regarding mathematics and statistics.

The evaluation of the layout of multidimensional objects in the n-dimensional space when $n > 3$ is impossible. This problem may be solved using multidimensional object projections into 2—or 3–dimensional space, which are more common for everyone. The use of different projections leads to different results. There are cases when projections do not present the actual data structure: clusters, outliers, and the layout of data objects.

Projection methods, called dimension reduction techniques, map multidimensional objects into lower-dimension space. The purpose – is to map objects from multidimensional space into lower dimensional space, keeping the data structure the same as in multidimensional space. Multidimensional space is projected into R^d–dimensional space, where $d = 2$ or $d = 3$.

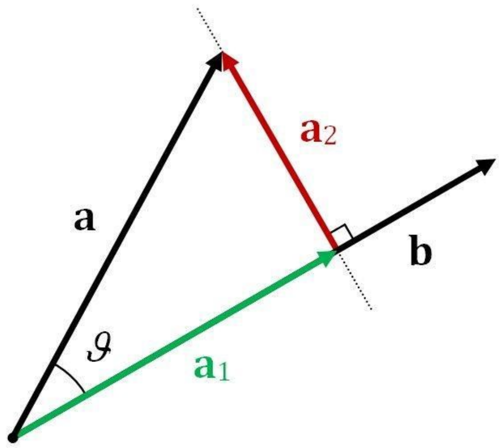

Figure 4.27: Projection illustration.

Let us have data represented by a matrix:

$$X = \{X_1, X_2, \dots, X_n\} = \{x_{ij}, i = 1, \dots, m, j = 1, \dots, n\},$$

which i^{th} row is data object defined by vector $X_i \in R^n$, where $X_i = (x_{i1}, x_{i2}, \dots, x_{in}), i \in \{1, \dots, m\}$ and x_{ij} is j^{th} attribute of ith object. The purpose – to find the transformation of the vector $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ into $Y_i = (y_{i1}, y_{i2}, \dots, y_{in})$ which is in lower dimensional projection space $R^d, d < n$.

- **Linear projection**. Linear projection is described by a linear equation system:

$$Y_i = X_i A.$$

Common case: $Y_i = (y_{i1}, y_{i2}, \dots, y_{in})$, $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, A – quadratic matrix $n \times n$. If linear transformation is used for dimension reduction, then $Y_i = (y_{i1}, y_{i2}, \dots, y_{id}), d < n, A - n \times d$ matrix.

- **Nonlinear projection**. Nonlinear projection is described by the function:

$$Y = f(X),$$

where f is nonlinear transformation.

Further, this section presents two-dimensional reduction methods: (1) linear – Principal components analysis and (2) non-linear – multidimensional scaling.

4.3.1. Principal Components Analysis

Principal component analysis (PCA) algorithm:

1. Let's look at

$$X = \{X_1, X_2, \dots, X_n\} = \{x_{ij},\ i = 1, \dots, m,\ j = 1, \dots, n\},$$

where the ith row is a data object defined by vector $X_i \in R^n$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i \in \{1, \ldots, m\}$ and x_{ij} is j_{th} attribute of i_{th} object.

2. Calculate correlation coefficients r_{kl} and correlation matrix $R = \{r_{kl}, kl = 1, \ldots, n\}$. Diagonal elements of the matrix R are equal to 1.

3. Calculate covariation coefficient c_{kl} and covariation matrix $C = \{c_{kl}, kl = 1, \ldots, n\}$. Matrix C – symmetric.

4. Covariation coefficients may express correlation coefficients:

$$r = \frac{c_{kl}}{\sqrt{c_{kk}c_{ll}}}.$$

If x_k and x_l are not correlated, their covariation coefficient is equal to 0: $c_{kl} = c_{lk} = 0, k \neq l$.

5. Define eigenvectors E_k and eigenvalues λ_k of the covariance matrix by solving equation $CE_k = \lambda_k E_k$. Here E_k is column vector, C – covariation matrix, λ_k– from the equation $|C - \lambda_k l| = 0$, where l – identity matrix with the same size as the matrix C, |•| defines determinant. The number of eigenvectors is equal to number of attributes n.

6. Sort eigenvectors E_k according to eigenvalues in descending order $(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n)$. Matrix $A = (E_1, E_2, \ldots, E_n)$ – matrix of the principal components. Its columns are eigenvectors $E_k, k = 1, \ldots, n$ which corresponds eigenvalues $(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n)$.

7. Perform transformation of the data vector $X_i, i = 1, \ldots, m$ using formula: $Y_i = (X_i - \mu)A$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), \mu = (\mu_1, \mu_2, \ldots, \mu_n)$ and $A = (E_1, E_2, \ldots, E_n)$.

8. Obtained vectors $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ represent points in the new orthogonal coordinate system which is defined by eigenvectors $E_k, k = 1, \ldots, n$. Covariation matrix of the components $y_1, y_2, \ldots, y_n$ of the vector $Y_i, i = 1, \ldots, m$ is:

$$(\lambda_1 \;\; 0 \;\; \ddots \;\; 0 \;\; \lambda_2)$$

9. Original data vector X_i may be extracted by $Y_i: X_i = Y_i A^T + \mu$.

10. Only several first eigenvectors are used for multidimensional data transformation. Let the first d eigenvectors compound matrix A_d. Then, the transformation: $Y_i = (X_i - \mu)A_d, i = 1, \ldots, m$. In that way, data vector projection into d– dimensional space is found ($d = 2$ or $d = 3$).

Figure 4.28 provides a visual representation of all the steps involved in representing the data in a less dimensional space, and Figure 4.29 provides an example of how this approach is used.

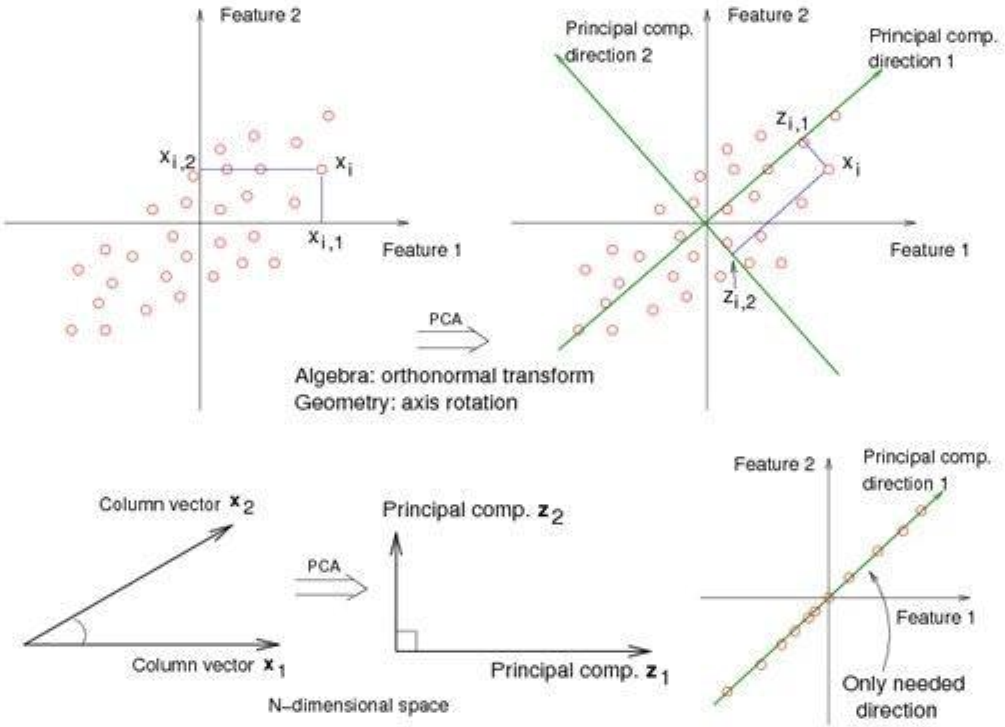

Figure 4.28: Principal components analysis algorithm visual representation.

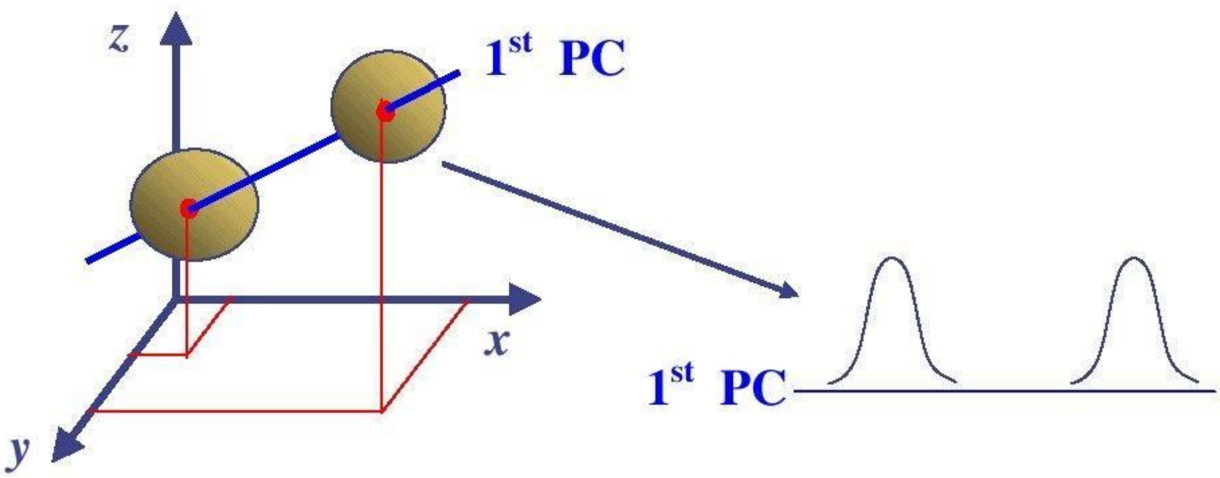

Figure 4.29: Principal components analysis example.

Principal components analysis has some limitations:

- Non-linear relations – impossible to identify.
- Outliers modify deviation (and covariation matrix) values, and “rotate” PC’s.

4.3.2 Multidimensional scaling

One powerful multidimensional data visualisation technique is multidimensional scaling (MDS) (Chen et al., 2007). MDS is widely used for exploratory data analysis in various fields, e.g., social sciences, economics, medicine, etc. The purpose of MDS is to provide a visual representation of similarities or distances between objects.

Let a set of n multidimensional vectors (representing the considered objects) have to be visualised in a $p = 2$ dimensional space. Pairwise dissimilarities measured between all pairs of objects are denoted by $\delta_{ij}, i, j = 1, \dots, n$. It is supposed that dissimilarities are symmetric $\delta_{ij} = \delta_{ji}$ and $\delta_{ii} = 0$. The points $x_1, x_2, \dots, x_n$ that constitute a set of n objects in p dimensional space should be found fitting pairwise distances of points to given pairwise dissimilarities $\delta_{ij}, i, j = 1, \dots, n$. The fitness criterion, called STRESS function should be minimized:

$$S(X) = \sum_{i=1}^{n} \sum_{j=1+1}^{n} w_{ij}\left(d(x_i, x_j) - \delta_{ij}\right)^2$$

where $X = (x_1, x_2, \dots, x_n), x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) d(x_i, x_j)$ denotes the distance between the points x_i and x_j and w_{ij} denotes weights.

The quality of multidimensional data projection into low-dimensional space depends on $d(x_i, x_j)$, and therefore the choice of metrics of the embedding space is essential.

Pairwise dissimilarities between pairs of objects may be considered as a distance in multidimensional original space and may be estimated using different norms in R^p. The most widely used distance measure is a Minkowski distance:

$$d(x_i, x_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right)^{\frac{1}{r}}$$

Parameter r influences the quality of projection into low-dimensional space. Usually, two well-known special cases of Minkowski distance are used in MDS: Euclidean distance when $r = 2$, and city-block distance when $r = 1$. The points $X = (x_1, x_2, \dots, x_n)$ found by means of minimisation of STRESS function using different distance metrics are different nonlinear projections of the set of objects in the original multidimensional space to the lower dimensional embedding space. Minimisation of STRESS function using Minkowski distance with various r (not only $r = 1$ and $r = 2$) may produce even more informative results.

The variety of solutions corresponding to $1 \leq r \leq 2$ may be ensured using a multiobjective optimisation approach. The multiobjective multidimensional scaling method, which may be helpful for exploratory multifaceted data analysis, is suggested in (Mackute-Varoneckiene et al., 2009).

An illustration of the multidimensional scaling method is presented in Figure 4.30.

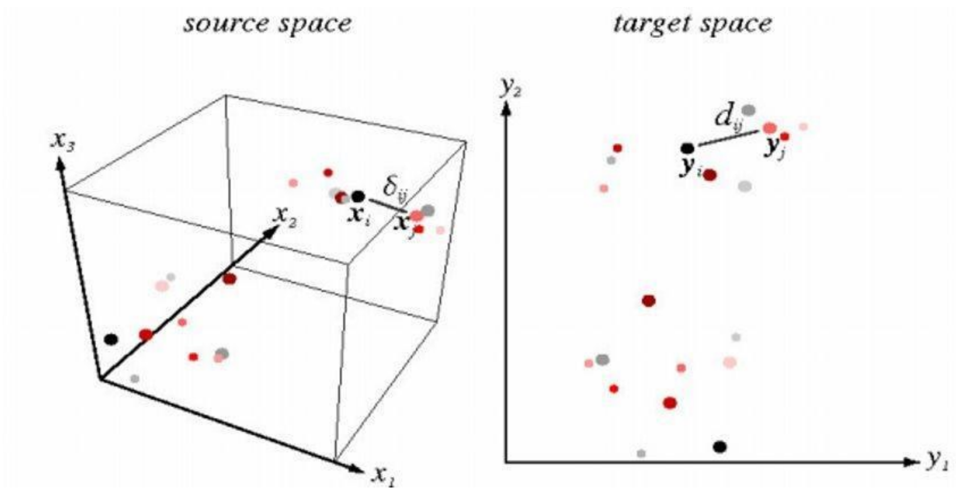

Figure 4.30: Multidimensional scaling illustration.

4.4 DATA VISUALISATION TOOLS/SOFTWARE

We've discussed various analyses and visualisation types; however, at times these visualisations might be more difficult for the average person with limited mathematical and statistical background to create or even to interpret. Therefore, a short webinar was created to better inform you about additional data visualisation tools such as Tableau, Google Charts, JupyteR, Dundas BI, and Infogram, as well as some other tools. To learn more about their tools, click here to watch the webinar by Ms Ilenia Chiodi from i-Strategies, Italy.

Chapter 5: Story-based Data Analytics

Prof. Haibo Li
Department of Media Technology and Interaction Design
KTH, Royal Institute of Technology, Sweden

TO WATCH THIS CHAPTER VIA WEBINARS, CLICK HERE

Data analytics can serve three pivotal information processing goals: surveying to summarise large datasets into understandable formats, confirming to test hypotheses or theories, and exploring to uncover hidden patterns and relationships. Data-driven and decision-oriented methods are two widely adopted approaches in data analytics, focusing on transforming raw data into actionable insights for decision-making. However, a fundamental limitation of these approaches is their linear and unidirectional nature, often inadequate for thoroughly exploring information necessary to solve complex or 'wicked' problems.

To overcome the limitation, we introduce a novel approach named "story-driven data analytics." This approach is notable for its focus on storytelling and collaboration between technical and business teams. Unlike the other approaches, it is iterative and flexible, beginning with a broad analysis and gradually becoming more focused. This characteristic makes it particularly effective for exploring information and dealing with wicked problems. Data analysis within a narrative framework enhances understanding and cooperation across various domains. This approach is ideal for complex problem-solving situations requiring diverse perspectives and expertise.

5.1 Introduction

We recently initiated a collaborative data analytics project with a renowned European wine company. The project aims to explore *whether Big Data analytics can revolutionise the century-old wine industry*. Undertaking the study raises a primary question: 'How should one conduct a data analytics project?' This proves challenging, as data analytics remains largely unexplored territory for many in the wine industry. There's uncertainty about the potential of Big Data analytics and, more importantly, a lack of clarity regarding their specific goals in this new field. This scenario illustrates typical challenges faced in diverse business contexts when initialising a data analytics project.

To start such a project, it's advisable first to gain a comprehensive understanding of the Big Data stack. Why is this crucial? Understanding the Big Data stack provides a roadmap for executing a data analytics project (Figure 5.1). Delving into the stack, we find the infrastructure layer at the base, encompassing tools, technology, personnel, and processes. Sitting directly above is the data layer, which can be categorised into two types: primary and secondary. Secondary data is obtained from existing datasets, while primary data is gathered directly through the firsthand collection. Sandwiched between the data and insight layers is the analytics layer, which houses the algorithms that form the backbone of data analytics. At the top of the stack, the focus is on the business goal, which can be achieved through three primary objectives: acquiring business information, validating beliefs, and conducting exploratory analysis.

With a thorough understanding of the Big Data stack, we can effectively use it to guide the execution of a data analytics project, ensuring each step, from goal setting to data selection and algorithm application, is strategic and informed.

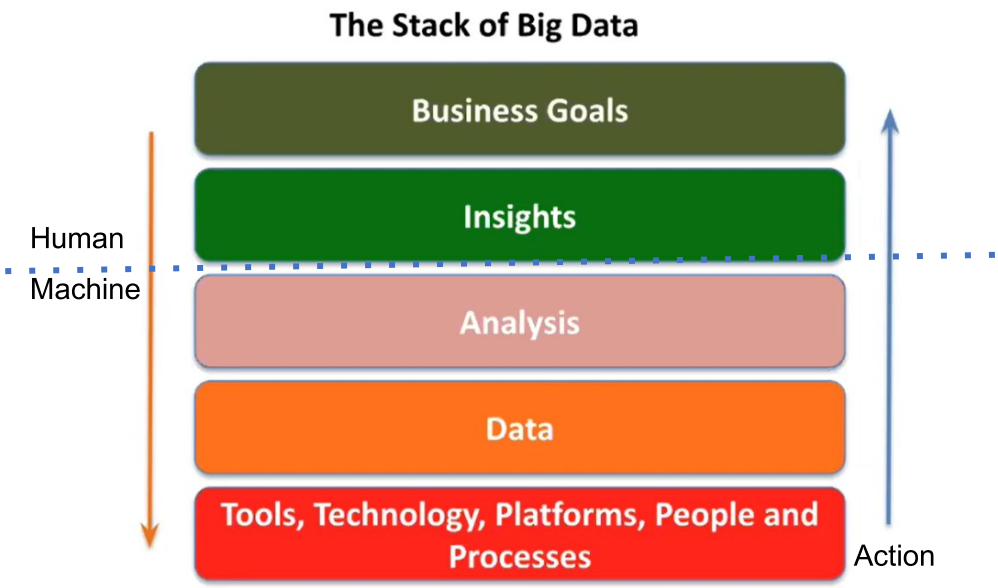

Figure 5.1: The Big Data Stack.

5.2 A Creative Data-driven Approach

Data-driven data analytics refers to an approach in analytics that emphasises the critical role of data in the decision-making process. In this approach, decisions are based not on intuition or personal experience alone but primarily on insights derived from data analysis.

From the data, we can directly undertake four types of analytics: *descriptive, diagnostic, predictive,* and *prescriptive* (Figure 5.2). These can be conducted using machine learning algorithms like supervised, unsupervised, and reinforcement learning.

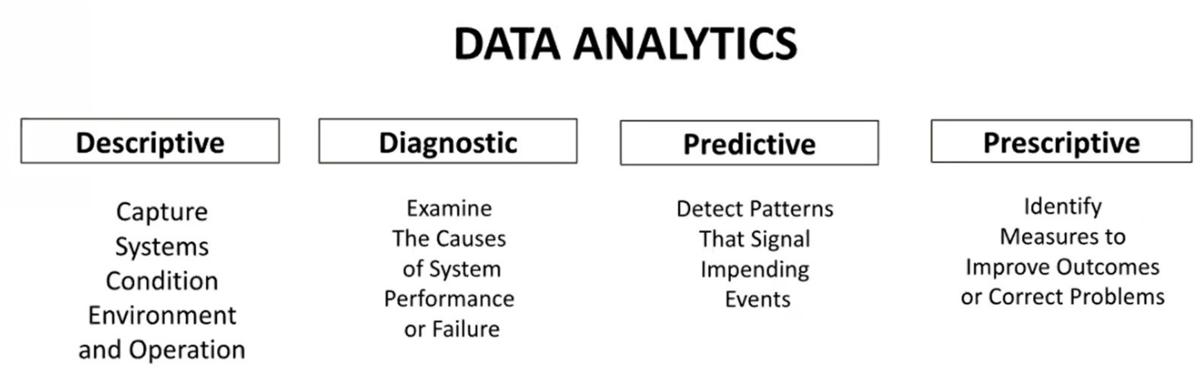

Figure 5.2: Data analytics

Data analytics fundamentally enables us to answer five key questions, all rooted in data: What happened? What is happening? Why did it happen? What will happen? and What actions should be taken? In this data analytics framework, data scientists can explore various datasets (Table 5.1). The objective is to stimulate their creativity and enable them to unearth unexpected and valuable insights, patterns, or connections within the data.

Table 5.1: Data-Mining Algorithms for finding patterns in data.

Tasks	Classification	Regression	Clustering	Association	Sequential association analysis	Anomaly detection	Text mining
(Patterns) Description	Sorts data into one of several predefined classes	Finding a function that fits the data	Finds groups of objects (clusters)	Discover association rules between frequent item sets	Search for relationships between variables	Identifies unusual data records	Derives patterns and trends from text
Algorithms	NBC, SVM, NN	SVM, CNN	K-mean	Aprior	HMM	RF, SVM	LLM, NBC

A classic example from data analytics is the discovery of an unexpected correlation between beer and diaper sales through sales data analysis. Similarly, in the wine industry, analysts might find an intriguing correlation between wine sales and airline ticket purchases. This could imply that incorporating tourist data might enhance wine sales analytics if valid. However, it's crucial to acknowledge the significant risk of overreaching (see Figure 5.3) with creativity in a data-driven approach. This is similar to the 'Knowledge vs. Experience' image, where perceived patterns may not exist. For instance, seeing a cat in this image would be a clear case of overreach or, in machine learning terms, '*overfitting*'.

An even more striking example would be claiming a correlation between wine sales and US crude oil imports from Norway after data analysis. Such an assertion would likely represent a spurious relationship rather than a meaningful correlation. De Langhe and Puntoni (2020) argue that, in practice, making decisions with data often comes down to finding a purpose for the data at hand. Companies look for ways to extract value from available data, but that doesn't necessarily mean data analysts are answering the right questions. It's also not a safeguard against the influence of preexisting beliefs and incentives.

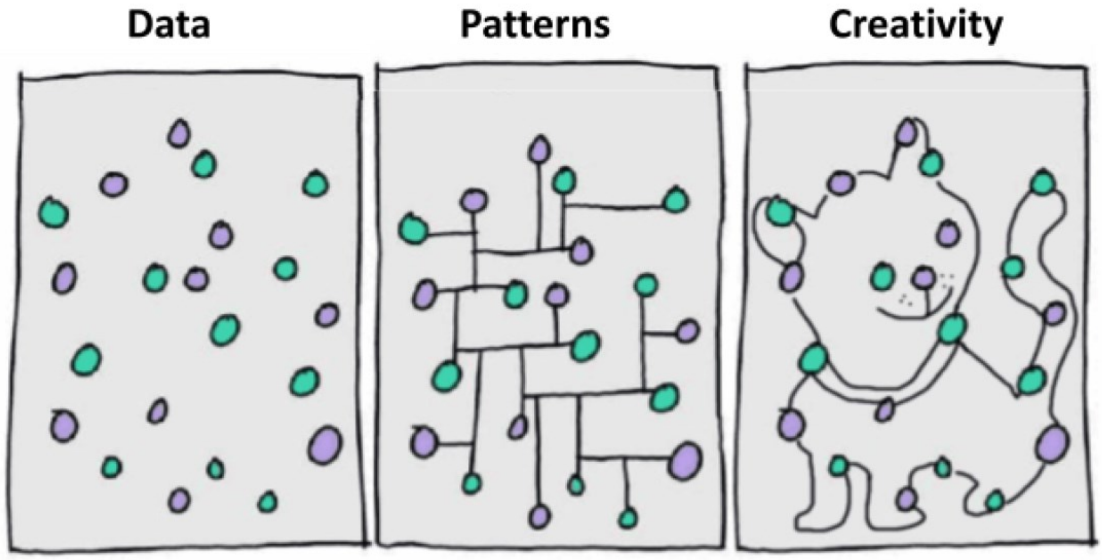

Figure 5.3: Knowledge vs. Experience Image (https://twitter.com/letkeman/status/423185090890321920).

De Langhe and Puntoni (2020) state that the solution is simple: *instead of finding a purpose for data, find data for a purpose*. They call this approach *decision-oriented data analytics*.

5.3 Leading with Decision-driven Data Analytics

Decision-oriented is a top-down approach, starting data analytics from business goals. In many companies, managers are 'hesitant to allow technical experts to conduct data analytics without specific goals or direction'. They often complain that the insights derived from such analytics are irrelevant. These managers, believing they have extensive experience and knowledge in the field, usually have a clear vision of the types of associations they seek. For example, they might assert that the quality of wine is the top priority and are firmly convinced of a relationship between wine taste and weather. They want to use the weather data to predict the quality of wines. They direct the technical team to identify such a relationship using data analytics. If the team doesn't have access to weather data, the company is willing to 'purchase the necessary datasets'. The data scientist's role is to compile data to support their findings.

This illustrates the advantages of a decision-oriented approach: the results are relevant and actionable because the managers give the tasks, and it is easy to get resources if needed. Since it is a top-down approach, whether successful depends on how people set down their business goals.

Let's take a closer look at the three primary objectives of setting business goals in the context of data analytics: *acquiring business information*, *validating beliefs*, and *performing exploratory analysis*. This is highly related to how people process information:

- **Information Survey**

Quite often, it is the primary objective of data analytics. These include gathering business-related information, such as surveying current business or sales data.

- **Information Confirmation**

Another purpose is to validate certain beliefs; for instance, a manager might firmly believe that the wine sale depends on inflation and needs data to substantiate his judgment.

- **Information Exploratory**

Data exploration is needed when situations are complicated, for example, when we don't know where we are going or how to get there. Such problems are called *wicked problems*. Here, we see some characteristics of a wicked problem. For such a complex issue, we must employ data analytics for exploitation, both to define the problem and to identify solutions.

Armed with clear business objectives, we can extract insights from the data, aligning with the Big Data stack's second layer. Insights are typically gleaned through four types of data analytics: *descriptive, diagnostic, predictive*, and *prescriptive*. Collectively, these approaches enable us to answer critical questions and construct distinct data analytics narratives.

The primary drawbacks of the decision-oriented approach include a significant reduction in the *creativity* of data scientists and, secondly, an increased risk of relying on data to support decisions, leading to misleading outcomes. De Langhe and Puntoni (2020) provide some tips on avoiding these problems and running successful decision-oriented analytics.

The top-down approach emphasises starting from business goals. Managers with a clear vision direct technical teams to identify specific relationships using data analytics. However, this approach carries a potential ethical risk: it might unintentionally mislead stakeholders. The decision-oriented approach works best when people look for information surveys or confirm their beliefs but will not work for information exploration. In the following section, we suggest a new story-driven data analytics approach.

5.4 STORY-DRIVEN APPROACH TO INFORMATION EXPLORATION

We observe that both data-driven and decision-oriented approaches have their advantages and disadvantages. A common issue in data analytics with these approaches is that they follow a *unidirectional* process, either bottom-up or top-down. These approaches are only effective for '*tamed*' problems where the issues can be clearly defined. However, most real-world data analytics tasks involve '*wicked*' problems, which are far more complicated. Let us have a close look at the wicked problem.

5.4.1 THE WICKED PROBLEM

A "wicked problem" is a term used to describe a complex issue that is difficult to define and inherently unsolvable. These problems are often social or cultural and are characterised by several distinct features:

- **Complexity:** Wicked problems are usually multifaceted, involving many interconnected factors, which makes them difficult to understand and address.
- **No Clear Solution:** Unlike "tame" problems with straightforward solutions, wicked problems have no definitive or universal answers. Solutions to wicked problems often lead to unforeseen consequences and new issues.
- **Stakeholder Diversity:** There are often many stakeholders involved, each with their own perspectives, values, and interests. This diversity can lead to conflicting viewpoints on how the problem should be approached or solved.
- **Evolving Nature:** Wicked problems are dynamic, meaning they evolve. What might seem like a solution at one point can become obsolete or inadequate as the problem changes.
- **Social Complexity:** These problems are typically embedded in social contexts and involve human behaviour, making them inherently complex and unpredictable.
- **No "True" or "False" Solutions:** Solutions to wicked problems are not right or wrong in the traditional sense but are often judged as "better," "worse," "good enough," or "not good enough."
- **Interconnectedness:** Wicked problems are often a symptom of another problem. This interconnectedness means addressing one aspect of a wicked problem can affect others.

Big Data analytics could offer innovative approaches to tackle the wicked problems in three aspects:

i. **Defining the Problem with Data Analytics**

- Complex Nature of Wicked Problems: Big Data analytics aids in unravelling the various dimensions of wicked problems by analysing data from multiple sources, thus providing a comprehensive understanding.
- Identifying Patterns and Trends: Analytics reveal patterns, trends, and correlations in large datasets, offering insights into underlying causes and effects within the wicked problem.

ii. **Iterative Exploration for Solutions**

- Adaptive Approach: The evolving nature of wicked problems necessitates flexible and adaptive solutions facilitated by continuous monitoring and analysis through Big Data analytics.
- Experimentation and Learning: Analytics supports testing various hypotheses and strategies, enabling iterative learning and refinement of approaches.

iii. **Collaborative Effort Between Business and Technical Teams**

- Diverse Perspectives: A cooperative team of business personnel and technical experts ensures a holistic understanding, combining practical implications and in-depth technical knowledge.
- Effective Communication and Decision Making: This collaboration enhances stakeholder communication, ensuring that data insights translate into actionable strategies and align with organisational goals.

5.4.2 Storytelling

Storytelling is crucial when tackling wicked problems with data analytics because it transforms complex, often abstract, data into a narrative that resonates with various stakeholders. This method is particularly effective because humans are naturally inclined to understand and remember stories better than raw data. Through storytelling, data analytics is not just presenting numbers and graphs; it's weaving a narrative that highlights cause and effect, making the problem more tangible and the need for action clearer.

Moreover, storytelling in data analytics facilitates better decision-making. A compelling story does more than just inform stakeholders; it also engages them emotionally, enhancing the impact of the data's implications. This emotional engagement is particularly critical in dealing with wicked problems, as these issues often require significant changes or interventions, which can be challenging to implement without strong stakeholder buy-in. Furthermore, a storytelling approach ensures that the insights gained from data analytics are accessible and compelling to a diverse audience, including policymakers, business leaders, and the public. It breaks down the barrier of technical jargon and complex statistical information, enabling a broader understanding and consensus, which is essential for addressing multifaceted wicked problems effectively.

In summary, storytelling in data analytics is not just about conveying information; it's about creating a narrative that motivates action and facilitates a deeper understanding of complex issues.

5.4.3 Data Analytics Driven by Story

Solving wicked problems with Big Data analytics is essential due to its role in complex decision-making, resource optimisation, and risk mitigation. It necessitates a blend of technical data analysis and practical business insights, enabling the development of innovative, adaptive solutions. The challenge is that managers and technical teams communicate in fundamentally different languages. It is imperative, therefore, for them to find common ground for communication. A narrative or 'story' can act as this unifying language. By employing storytelling, they can effectively navigate data analytics across two crucial dimensions: *problem focus* and *time framework*. The *focus* aspect allows for in-depth discussions about the '*what*', '*why*', and '*how*' of the problem. Meanwhile, the *time* framework guides decisions about the data types to be collected and analysed over time.

After establishing the type of story, storytellers and data scientists can re-engage with the technical team. In this collaboration, they delve into specific details, discussing the application of machine-learning techniques like clustering or regression. They also consider the nature of the data available—whether it's first-hand or second-hand—for the analytical aspects of the story. Thus, we can put all terms on a table over; business and technical people work interactively to explore data and tasks.

For instance, consider using advanced prescriptive data analytics to uncover actionable insights. Imagine a scenario centred on 'discounts', such as exploring whether a 10% discount on Saturdays could boost wine sales. To narrate this story, data analytics is employed to identify the strategic actions that should be taken. In contrast to traditional methods, a story-based approach focuses on storytelling and human-centric thinking. It requires collaboration between business and technical experts, using a 'story' as a common language to guide discussions about the problem focus and time frame in data analytics. This is a flexible, iterative method of data exploration led by a storyteller. It aims to reveal meaningful insights through a dynamic analysis process that begins with a broad scope and gradually becomes more focused. This method is distinctive for its ability to discover hidden narratives within the data, emphasising the unfolding of stories to understand and communicate complex information.

Table 5.2: Effective communication in the team

		Focus		
		What	Why	How
Time	Past	Q1: What happened?	Q4: Why happened?	
	Present	Q2: What is happening?		
	Future	Q3: What will happen?		Q5: How to improve?

Table 5.3: Technical communication

		Story Type				
		What happened?	What is happening?	What will happen?	Why it happened?	How to improve
Data Analytics	Descriptive	Clustering Classification		Regression		
	Diagnostic				Causal analysis	
	Predictive			Regression Neural Networks		
	Prescriptive					Reinforcement learning

5.5 Concluding Remarks

This chapter has introduced three approaches to the initial question of executing a data analytics project. Here is a summary of these three approaches.

Data-Driven Approach: This approach centres around decisions and insights derived from data analysis. The creativity of data analysts is a key advantage, allowing them to uncover unique insights and novel solutions. However, a potential risk in this approach is the generation of spurious relationships, where data is misinterpreted, or misleading connections are made. Despite this, the emphasis remains on leveraging empirical evidence and observations to guide actions.

Decision-Oriented Approach: Contrary to the data-driven approach, the decision-oriented approach starts with managers' personal beliefs or hypotheses. This method seeks data to support these pre-existing beliefs or decisions. While this can streamline the focus of data collection and analysis, it may raise ethical concerns, as the data used could be selectively chosen to validate a specific viewpoint, potentially ignoring other crucial data that might contradict or challenge these beliefs.

Story-Based Data Analytics: This approach is characterised by its emphasis on storytelling and collaboration between technical and business teams. It's an iterative and flexible method that starts with a broad analysis and progressively narrows down. This approach is particularly effective for exploring information and addressing wicked problems. Framing data analysis within a narrative facilitates better understanding and cooperation across different domains, making it an ideal strategy for complex problem-solving where various perspectives and expertise are essential.

Table 5.4: Managing a Data Analytics Project

	Data-driven	Decision-driven	Story-driven
Purpose	Information	Confirmation	Exploration
Led by	Data Scientists	Managers	Storyteller
Data	Utilising the existing data	Looking for new data	Starting with the existing data, new data will be added when necessary
Scope	Wide	Narrow	Wide first narrow, then
Data Analytics	One-way process bottom-up	One-way process top-down	Interactive process exploration
Creativity	Data scientist	Manager	The team (Storyteller)

Figure 4: Example Image (Source: Created by DALL-E 2, an AI system designed by OpenAI.)

A comparative analysis of these methods, as shown in Table 3, will assist in determining the most suitable approach for specific cases. In the final remark, we'd like to highlight the impact of ChatGPT, especially with its robust data analytics plugin. ChatGPT introduces an intuitive approach to data analytics, where natural language precedes traditional programming languages. This shift is particularly evident in the use of prompt engineering. By crafting prompts, we pave the way for a future in data analytics where data scientists, managers, and storytellers collaborate closely, supported robustly by computers. In this scenario, the prompt engineer plays a pivotal role as the storyteller, guiding the team through the storytelling process with ease and finesse. This innovative approach marks a significant shift in how we approach data analysis, blending technical expertise with the art of storytelling.

We see a significant development in data analytics with the introduction of ChatGPT, which prioritises intuitive, natural language interactions over traditional programming languages. This shift is especially apparent in the practice of "prompt engineering." By carefully designing prompts, a pathway is forged toward a future in data analytics where there is close collaboration among data scientists, managers, and storytellers, all supported effectively by advanced computing technologies.

Chapter 6: The Case of Big Data in intangible cultural heritage tourism (TM)

Dr Marco Scholtz
Ms Kaat De Ridder
Center for Expertise for Sustainable Business and Digital Innovation
Thomas More University of Applied Sciences, Mechelen, Belgium

TO WATCH THIS CHAPTER VIA WEBINARS, CLICK HERE

This chapter examines the intersection of Intangible Cultural Heritage (ICH), tourism, and Big Data, highlighting their collective impact on the tourism industry. It's divided into three sections: the first explores the significance of heritage tourism, especially ICH, and its socio-economic impacts, including the 2003 Convention for Safeguarding ICH. The second section discusses the role of Big Data in tourism, addressing current data limitations and illustrating its potential through a case study on the UNWTO's use of Big Data for storytelling. The final section presents practical applications of Big Data in enhancing ICH tourism storytelling, with an example involving a UNESCO ICH dataset. This chapter aims to equip students and educators with a comprehensive understanding of leveraging Big Data to improve ICH tourism experiences, fostering sustainable and meaningful tourism practices.

6.1 INTANGIBLE CULTURAL HERITAGE AND TOURISM

Tourism is a vital industry with far-reaching economic and social impacts, contributing significantly to global GDP and job creation (WTTC, 2023). Heritage tourism, encompassing tangible and intangible aspects, plays a crucial role in this sector. Intangible Cultural Heritage (ICH) offers enriching experiences deeply rooted in local identity and social tissue and has gained prominence in cultural heritage tourism. The 2003 UNESCO Convention provides a framework for safeguarding ICH and demonstrates its value in enriching tourist experiences (UNESCO, 2022).

This section aims to explore the importance of tourism and heritage, focusing on the role and value of ICH. We will also discuss how the 2003 Convention supports the tourism industry in preserving these invaluable cultural assets.

6.1.1 Why is the tourism industry important?

Tourism significantly contributes to global economies, fostering economic growth, cultural exchange, and, in some well-managed instances, environmental conservation. This section explores the multifaceted importance of the tourism industry, drawing insights from reputable sources.

- **Economic Significance**

The economic impact of tourism is substantial. According to the World Travel & Tourism Council (WTTC), the tourism sector accounted for 7.6% of global GDP in 2022 (WTTC, 2023). This contribution extends across various sectors, including hospitality, transportation, and entertainment. The tourism industry generates employment opportunities, particularly in developing regions, as highlighted by the United Nations World Tourism Organization (UNWTO, 2023). This economic boost plays a pivotal role in poverty alleviation and community development.

- **Cultural Exchange and Global Understanding**

Tourism facilitates cultural exchange and fosters global understanding. As noted in an academic article by Hall and Lew (2009), tourism promotes intercultural encounters, which can reduce prejudices and stereotypes among travellers and host communities. This exchange of ideas and traditions leads to cultural enrichment and mutual respect.

- **Environmental Conservation and Sustainability**

Sustainable tourism practices are increasingly emphasised for their role in environmental conservation. The United Nations Environment Programme (UNEP, 2021) highlights that responsible tourism can contribute to the protection of natural resources and ecosystems. By promoting eco-friendly initiatives and minimising negative environmental impacts, the tourism industry can play a vital part in safeguarding the planet.

In conclusion, the importance of the tourism industry transcends economic benefits. It catalyses economic growth, cultural exchange, and environmental conservation. These insights, supported by sources such as the WTTC, UNWTO, academic articles like Hall and Lew (2009), and UNEP reports, underscore the industry's pivotal role in shaping our global society and economy.

6.1.2 Intangible Cultural Heritage (ICH) Tourism

We all understand the term' heritage', but do we? According to Erfgoedcellen.be (n.d.), ‘heritage’ encompasses the legacies passed down from our ancestors that we choose to maintain and honour. This includes physical structures and monuments, often called 'immovable heritage', and portable historical items like artefacts, artworks, and relics, which constitute 'movable heritage'. Beyond these physical items, heritage also consists of more transient elements like oral traditions, festivities, music, languages, and craftsmanship, collectively known as 'intangible heritage' (See Figure 6.1 for comparisons).

Immovable tangible heritage | Moveable tangible heritage | Intangible cultural heritage

Figure 6.1: Illustration of the different forms of cultural heritage (Source: Created by DALL-E 2, an AI system designed by OpenAI.)

Cultural Heritage Tourism (CHT) consists of three different forms of heritage. However, it's been shown that in most instances, the tangible forms of heritage enjoy the most attention, while intangible cultural heritage (ICH) often receives less attention (Doesselaere, 2020). By actually incorporating ICH into CHT experiences, it can provide more depth and context to heritage visitor experiences (Kirshenblatt-Gimblett, 2014), which can lead to various economic and social benefits such as increased visitor satisfaction, memorable experiences, positive word-of-mouth, repeat visits, improved safeguarding[10] of ICH, enhanced educational value and evocation of emotional responses (Petronela, 2016; Masoud, Mortazavi, & Frasani, 2019).

There has been a push towards strengthening a thoughtful and sustainable partnership between Intangible Cultural Heritage (ICH) and Cultural Heritage Tourism (CHT) experiences

[10] Safeguarding: Safeguarding is different to preservation as it does not 'fix' or 'freeze' ICH – it is about the transferring of knowledge, skills and meaning that will continuously evolve due to changing human behaviour and technological advancements (UNESCO, 2009)

to achieve the previously outlined benefits. This occurs within various cultural heritage and tourism organisations, academic institutions, and even through UNESCO (2011).

6.1.3 UNESCO's 2003 Convention for ICH

UNESCO, the acronym for the United Nations Educational, Scientific and Cultural Organisation, fosters global peace and security through collaborative efforts in education, science, culture, communication, and information. It advocates for exchanging knowledge and the unimpeded circulation of ideas, facilitating deeper mutual comprehension and a fuller appreciation of diverse ways of life. The initiatives undertaken by UNESCO are instrumental in fulfilling the Sustainable Development Goals outlined in the 2030 Agenda, which the UN General Assembly endorsed in 2015 (UNESCO).

In 2003, a milestone was achieved when UNESCO decided to put together the Convention for the Safeguarding of ICH. This is a milestone as UNESCO finally indicated the importance of ICH and safeguarding practices surrounding it. This Convention has been updated over the years. Currently, it contains essential information such as 1) a representative list of ICH of humanity, 2) a register of Good Safeguarding Practices, and 3) an urgent safeguarding list (UNESCO, 2022). The tools and data obtained from this Convention play a crucial role in understanding ICH worldwide and can be used by tourism and cultural heritage stakeholders, policymakers, and everyday citizens better to understand the current scope of ICH and its safeguarding. We will focus on this data later in this chapter.

6.2 The use of data in tourism

Tourism generates a significant global economic impact, and understanding its dynamics through data is crucial for sustainable development. Traditional tourism datasets have limitations in providing comprehensive insights into this complex industry. This section explores the significance of tourism data and the challenges posed by current datasets. It then introduces Big Data as a promising avenue for enhancing our understanding of tourism and discusses its potential benefits and users. Additionally, we delve into the underutilised aspect of Big Data storytelling in the tourism context. Finally, we examine a real-world case study highlighting the United Nations World Tourism Organization's (UNWTO) use of Big Data for impactful storytelling, demonstrating the power of data-driven narratives in tourism.

6.2.1 The significance of tourism data

Tourism data plays a pivotal role in understanding and shaping the global travel industry. It represents a vast collection of information generated by travellers, businesses, and organisations, offering valuable insights into the sector's dynamics. In this chapter, we delve into the significance of tourism data, shedding light on what it encompasses and why it is essential for various stakeholders.

6.2.1.1 Understanding Tourism Data

Tourism data encompasses various information related to travel and tourism activities. It includes data on tourist arrivals and departures, accommodation bookings, transportation, visitor demographics, expenditure patterns, destination preferences, and more. This data is collected through various sources, such as government agencies, tourism boards, hotels, airlines, travel agencies, and online platforms. The sheer volume and diversity of tourism data make it a valuable resource for decision-makers, researchers, and businesses in the tourism industry.

6.2.1.2 Why tourism data is needed

Data plays a pivotal role in comprehending and managing tourism (Grepsr, 2011; Mariani & Baggio, 2022; Mountasser, Ouhbi, Frikh, & Hdioud, 2020; Pangea X, 2022). Here are the critical roles of data in tourism:

- **Informed Decision-Making:** Tourism data gives decision-makers insights to formulate policies, strategies, and marketing campaigns. Understanding visitor trends and preferences enables destinations and businesses to enhance the visitor experience.
- **Economic Impact Assessment:** Tourism significantly contributes to global economies. Data allows governments and organisations to assess its economic impact, including contributions to GDP, job creation, and revenue generation.
- **Market Research:** Tourism data assists businesses in market research, helping identify target markets, consumer preferences, and emerging trends. This enables companies to tailor products and services to meet travellers' demands.
- **Infrastructure Planning:** Data informs infrastructure planning, including airports, roads, accommodation facilities, and attractions. Understanding visitor flows and transportation needs is vital for sustainable tourism development.

- **Competitive Advantage:** Tourism businesses gain a competitive edge by leveraging data to stay ahead of market trends, optimise pricing strategies, and target specific customer segments.
- **Sustainability and Environmental Management:** In the era of sustainable tourism, data is essential for monitoring and managing the environmental impact of tourism. It identifies areas of concern and supports eco-friendly practices.
- **Visitor Safety and Security:** Tourism data enhances visitor safety by tracking traveller movements, responding to emergencies, and implementing measures for a secure tourism environment.
- **Crisis Management:** During crises like natural disasters or health emergencies (e.g., the 2020 COVID-19 pandemic), tourism data is crucial for effective crisis management. It assesses disruptions and aids in planning recovery efforts.

In essence, tourism data serves as the foundation of the tourism industry. Its multifaceted role empowers stakeholders to make informed decisions, foster economic growth, promote sustainability, and create memorable experiences for travellers. As we delve into Big Data in tourism, we'll explore how harnessing this data can revolutionise storytelling and optimise the entire tourism ecosystem, resulting in more engaging narratives for travellers (demand) and destinations (supply).

6.2.2 Limitations of Traditional Tourism Datasets

Conventional tourism data sets, pivotal in analysing travel trends and patterns, encounter notable limitations in the burgeoning Big Data era. This section elucidates these limitations within the rapidly evolving tourism sector.

- **Lack of Real-Time Insights:** Conventional data sets, typically gathered monthly or yearly, do not adequately reflect the immediate state of the industry. In a sector where travel plans are subject to swift changes, this reliance on historical data curtails the ability to forecast future trends and respond to present demands.
- **Insufficient Depth of Information:** These data sets generally provide basic details such as the number of tourists and length of stay, yet they fall short of comprehensively understanding tourist behaviour and preferences. This leads to somewhat superficial profiling of visitors, lacking the nuanced insights necessary for devising personalised experiences, effective marketing strategies, and targeted infrastructure developments.
- **Restricted Geographical Scope:** Traditional data sets often concentrate on regional or national trends, missing out on the intricacies of international travel (Volo, 2020).

Conversely, Big Data extends its reach globally, offering more profound insights into the behaviours and preferences of global travellers (Yallop & Seraphin, 2020).

- **Challenges in Data Integration:** Conventional data sets are typically collated and stored in isolation, making amalgamating information from diverse sources challenging. This leads to a piecemeal perception of the tourism landscape.

In summary, while conventional tourism data sets are of value, their limitations are pronounced in today's data-driven tourism environment. The advent of Big Data brings new possibilities for obtaining dynamic, real-time, and detailed insights, thereby enriching tourism and cultural heritage storytelling.

6.3 Big Data in Tourism

In the preceding chapters, we have explored Big Data's vast and varied nature, its rapid accumulation, and the sophisticated methods required for its analysis. Key to multiple sectors like tourism and healthcare, Big Data provides invaluable insights for decision-making and strategic planning. Tourism enables organisations to enhance operations and customer service, ensuring competitiveness (Mariani et al., 2018; Buhalis and Sinarta, 2019). The essence of Big Data lies in managing large, diverse datasets swiftly for real-time analysis, often facilitated by AI and machine learning. The upcoming sections will delve into its application in tourism, its benefits to the industry, ways to share such data, and how storytelling can render it more accessible and engaging.

6.3.1 The use of Big Data in Tourism

Big Data can help tourism businesses and cultural organisations to (1) obtain a competitive advantage over their competitors while (2) building customer loyalty. It can do this by creating (Epamanywhere, 2022):

1. Improved, integrated[11] current understanding of consumers

- **Consumer Behaviour Analysis:** Tracks and analyses tourists' behaviours, preferences, and spending habits in real-time.

[11] Big Data refers to provided integrated data from various sources. Instead of interpreting various sets of traditional data, it brings various forms and sources of data together.

- Personalisation: Facilitates customised offerings based on individual tourist preferences.
- **Demand Forecasting:** Predicts future tourism trends and demands based on current data patterns.
- **Experience Enhancement:** Improves the overall tourist experience by identifying popular attractions and activities.
- **Market Segmentation:** Segregates tourists into different groups for targeted marketing and service delivery.
- **Feedback Analysis:** Utilises online reviews and social media data for service improvement.
- **Real-Time Insights:** Provides up-to-date information on tourist movements and preferences.
- **Competitive Analysis:** Helps understand competitor strategies and market positioning.

2. Swift identification of changes in industry patterns/trends

- **Disruptive innovations:** It can provide instant insights into how consumers use disruptive innovations, such as the sharing economy (e.g. Airbnb, Uber, Lift, etc.)
- **Health, natural, and political emergencies:** It can help us immediately understand the impact of events such as COVID-19, earthquakes, etc., on travel consumers.
- **(Immersive) technologies:** We can immediately see the impact of the latest technologies on consumers and their required experiences.

3. Fast, calculated industry predictions

- **Future-proof planning**: Already adapt products and services for future market dynamics (fast-changing markets and preferences)
- **Determine growth areas: It provides** a tool to quickly determine what parts of tourism or a company might grow, which allows for strategic planning to take advantage of such growth areas.

- **Improved revenue management**: The ability to see which revenue/income streams might be most relevant and expand those streams.
- **Seasonal management**: Determining what the market will look like or what to do during seasonality allows one to plan ahead and mitigate/minimise any negative impacts.

4. Improving accuracy of decision-making

Examples:

- **Revenue management:** Predict an event's profit based on the predicted attendee numbers or determine a hotel's occupancy rate in the following season to determine income levels.
- **Reputation management:** Analyse consumer interactions on location and online to identify businesses' strengths and weaknesses and adapt accordingly.
- **Targeting marketing:** Big Data allows the software to better understand consumers' preferences and needs and can automatically personalise marketing messages, deals, and pricing.
- **Enhanced customer experiences:** Travel agencies, accommodation, and car hire agencies can provide personalised amenities to their guests, which helps guarantee consumer satisfaction.
- **Strategic management:** Tourism organisations, businesses and entrepreneurs can use Big Data to do long-term planning (strategic planning), allowing for sudden market changes.

6.3.2 The kinds of Big Data collected in Tourism

As previously mentioned, Big Data allows for capturing various forms of data to help us understand markets, identify changes/patterns, make fast predictions, and make calculated, accurate business decisions. However, what types of tourism data can be collected? The following (Octoparse, 2021; Optimizely, 2022) are the most commonly captured data:

- **Consumer demographics and preferences:** Who are the consumers, where do they come from, and what are their preferences before, during, and after travels?
- **Tourist flows:** Number of people arriving and departing a destination, place of origin, etc.

- **Destination information:** Places to visit; rules and regulations; (intangible) cultural heritage on offer.
- **Events data:** All information regarding sports events, festivals, etc.
- **Hotel and restaurant data:** Capturing customer reviews, ratings, recommendations, prices, and service details (e.g. TripAdvisor, Google Reviews).
- **Transportation:** Air tickets, flight prices, traffic data, for example.
- **Booking activities:** Determine who is booking where and control prices according to demand and supply.

There are other possible kinds of data, but these are the primary types. You might wonder why some companies or businesses use such data; this will be discussed in the next section.

6.3.3 Businesses/organisations that make use of Big Data

Big Data is widely regarded as a valuable tool throughout the tourism industry due to the previously discussed benefits it can generate. To better contextualise it, here are examples of where Big Data is used. Table 6.1 overviews specific tourism stakeholders and how they use Big Data.

Table 6.1: An overview of tourism stakeholders who use Big Data.

Tourism stakeholder	Example	What it is used for	Source of information
Hotel chains	Marriot Hotels, Taj Hotel	**Customer experience enhancement:** Personalising guest experiences based on previous stays, preferences, and behaviours. **Dynamic pricing strategies:** Adjusting room rates in real-time based on demand, season, local events, and competitor pricing. **Operational efficiency:** Streamlining hotel operations such as housekeeping, maintenance, and staffing using predictive analytics. **Market and Trend Analysis:** To stay competitive, understand market trends, guest demographics, and emerging preferences. **Targeted marketing:** Creating personalised marketing campaigns and offers based on guest data and behaviour patterns. **Reputation management:** Monitoring and analysing online reviews and social media to improve service and address customer concerns. **Loyalty programs:** Enhancing loyalty programs by analysing member data to offer tailored rewards and experiences. **Risk Management and Security**: Identifying and mitigating risks through data analysis, including cybersecurity threats.	**Reservation systems:** Data from bookings, cancellations, and guest preferences gathered through online reservation platforms. **Loyalty programs:** Detailed customer information from loyalty program memberships, including stay history and preferences. **Social media and online reviews:** Analysing guest feedback and comments on social media platforms and review sites. **Point of Sale (POS) Systems:** Information from in-hotel purchases and services guests use. **Guest surveys and feedback forms:** Direct feedback from guests regarding their stay experience, amenities, and services. **Website and mobile app analytics:** Data on how guests interact with the hotel's digital platforms, including search behaviour and booking patterns. **Market research data:** Industry-wide data on travel trends, economic indicators, and competitor analysis. **Sensor and IoT data:** Use smart technology to gather guest preferences and behaviour data in rooms and facilities.

Restaurants	Uber Eats, Starbucks, McDonald's	**Customer Behaviour and Preferences:** Analysing data to understand customer preferences, peak dining times, and popular menu choices. **Menu optimisation:** Tailoring menu items based on customer preferences and sales data; also identifying trends for new dish introduction. **Inventory management:** Using predictive analytics for efficient stock management, reducing waste, and ensuring ingredient availability. **Operational efficiency:** Optimising kitchen operations, table management, and staffing based on data insights. **Personalised customer experience:** Creating personalised dining experiences based on customer history and preferences. **Dynamic pricing:** Implementing data-driven pricing strategies for special events, peak hours, or specific customer segments. **Location analysis:** Using demographic and geographic data to select new restaurant locations or market expansion strategies. **Marketing and promotion:** Targeted marketing campaigns and promotions based on customer data and dining trends.	**Point of Sale (POS) Systems:** Transactional data, including dish popularity, average spending, and customer return rate. **Customer feedback:** Online reviews, surveys, and feedback forms provide insights into customer satisfaction and areas for improvement. **Reservation and ordering systems:** Data from online reservations and ordering platforms provides insights into customer preferences and behaviours. **Social media analytics:** Understanding customer engagement and preferences through social media interactions and trends. **Supply chain data:** Information from suppliers for better inventory management and cost control. **Loyalty programmes:** Data on customer loyalty and habits from reward programs. **Market Research:** Industry trends, demographic data, and local market analysis for strategic planning. **Sensor and IoT Devices:** Using smart devices in restaurants to track customer flow, table turnover, and operational efficiency.
Airlines	United Airlines, EasyJet, Ryanair, Delta Airlines	**Customer preferences and behaviour analysis:** Understanding what customers prefer, their buying habits, peak times, and popular menu items. **Personalised marketing:** Using customer data to create targeted marketing campaigns and personalised offers. **Inventory management:** Predicting stock needs and managing inventory more efficiently based on consumption data. **Operational efficiency:** Analysing data for optimising kitchen operations, reducing wait times, and improving customer service.	**Point of Sale (POS) Systems:** Transactional data, including sales, time of purchase, and customer preferences. **Loyalty Programmes:** Data collected from loyalty card usage provides insights into customer behaviours and preferences. **Mobile Apps and Online Platforms:** Data gathered from customer interactions on mobile apps and websites, including ordering habits and preferences. **Social media and online reviews:** Analysing customer feedback and reviews on social media for trends and areas of improvement.

		Menu optimisation: Tailoring menu items based on regional preferences and seasonal trends identified through data analysis. **Dynamic pricing:** Implementing location-based or time-based pricing strategies based on demand and customer traffic. **Market expansion and location planning:** Using demographic and geographic data to decide where to open new outlets. **Customer feedback and improvement:** Gathering and analysing customer feedback for service and product improvement.	**Market research data:** Industry trends, demographic data, and competitor analysis to inform strategic decisions. **Sensor and IoT data:** Utilising store sensors for customer traffic patterns, seating preferences, and operational efficiency. **Supply chain data:** Supply chain logistics analysis for optimising inventory and reducing waste.
Sharing economy	Uber, Airbnb	**Market analysis and trends:** Big Data helps analyse market trends, understand customer preferences, and predict future demand. **Personalised customer experiences:** By analysing data, Airbnb can offer personalised recommendations to users, enhancing user experience. **Pricing strategies:** Dynamic pricing models are developed using Big Data to adjust prices based on demand, season, and other factors. **Risk management and security:** Data analysis helps identify potential fraud, ensure user verification, and maintain platform integrity. **Optimising operations:** Big Data improves operational efficiency by optimising property listings and enhancing customer service. **Marketing and advertising:** Tailoring marketing campaigns to specific demographics or user behaviours based on their data profiles. **Regulatory compliance and reporting:** Ensuring compliance with local laws and regulations through data analysis.	**User interactions:** Data collected from user interactions on the platform, like searches, bookings, reviews, and messages. **Social media:** Analysing social media trends and user feedback to understand public perception and demand. **Location data:** Geographic information from users and property listings to understand popular destinations and local market dynamics. **External market data:** To contextualise and forecast market conditions, economic indicators, tourism trends, and real estate market data. **User profiles:** Demographic and behavioural data from user profiles for targeted marketing and personalisation. **Transactional data:** Information from bookings, cancellations, and payments for financial analysis and strategy development. **Sensor data:** In some cases, IoT devices and property sensors can provide data on usage patterns and preferences.
Digital travel agencies	Cheapflights, Expedia, Booking.com, TripAdvisor,	**Personalised recommendations:** Offering tailored travel options based on past searches, bookings, and user preferences.	**User interaction data:** Data collected from user searches, bookings, clicks, and navigation patterns on their websites and apps.

	Skyscanner, Hopper, KindTraveler	**Dynamic pricing:** Utilising real-time data for pricing strategies, including flight and hotel rates, based on demand and availability. **Market trend analysis:** Analysing travel trends, popular destinations, seasonal demands, and emerging market opportunities. **Customer behaviour analysis:** Understanding user behaviour on their platforms to optimise user experience and increase conversions. **Predictive analytics:** Forecasting future travel trends, pricing patterns, and customer demands using historical and real-time data. **Optimised search functionality:** Enhancing search algorithms to provide users with more relevant and efficient search results. **Targeted marketing and promotions:** Creating targeted advertising campaigns and offers based on customer data and insights. **Risk management and fraud detection:** Data analysis identifies and mitigates potential risks and fraudulent activities.	**Customer reviews and feedback:** Insights from user reviews and feedback for improving services and addressing customer needs. **Social media engagement:** Analysing data from social media platforms to understand public perception and travel trends. **External data sources:** Collaborating with airlines, hotels, and other travel service providers for comprehensive data on availability and pricing. **Market research data:** Utilising industry reports, travel trends, and economic indicators for broader market insights. **Transactional data:** Information from bookings and transactions to understand customer preferences and spending patterns. **Loyalty programs and user accounts:** Data from user profiles and loyalty programs, including travel history and preferences.
Governmental organisations & Non-governmental organisations (NGOs)	United Nations World Tourism Organisation (UNWTO), Visit Flanders (tourism organisations), UNESCO	**Policymaking and planning:** Informing tourism policies and development strategies based on data-driven insights. **Monitoring and preservation of heritage sites:** Using data to monitor the condition of heritage sites and plan conservation efforts. **Sustainable tourism development:** Analysing data to promote sustainable tourism practices and manage environmental impacts. **Economic impact analysis:** Assessing the economic impact of tourism on local and national economies. **Crisis management and response:** Utilising data for effective response to tourism crises, such as natural disasters or pandemics.	**Tourist surveys and feedback:** Data from tourist surveys, feedback forms, and interviews. **Social media and online platforms:** Analysing social media trends, online reviews, and digital footprints of tourists. **Economic and statistical data:** Gathering data from financial reports, tourism spending, and employment statistics. **Environmental data:** Information related to environmental impact, biodiversity, and sustainability indicators. **Cultural and heritage site data:** Data from heritage site monitoring, including visitor numbers and site conditions.

		Visitor management and experience: Enhancing visitor experiences and managing visitor flows at popular sites to prevent overcrowding. **Cultural heritage and community engagement:** Understanding community needs and cultural dynamics to ensure inclusive and respectful tourism development. **Marketing and promotion:** Targeting and tailoring marketing campaigns for specific demographics and interests.	**Travel and hospitality industry data**: Collaborating with travel agencies, airlines, and hotels for comprehensive tourism data. **Governmental databases:** Utilising data from government databases on population, infrastructure, and regional development. **Geographic information systems (GIS):** Spatial data for mapping tourist flows, site locations, and infrastructure development.

This section underscores the challenges in interpreting Big Data in the tourism industry for the public, students, and professionals. Despite the abundance of data and advancements in Machine Learning and AI, comprehending the depth of Big Data remains complex. The crux of the issue is storytelling - transforming vast, intricate data into engaging and understandable narratives. Linking to an extensive library, Big Data requires skilled storytelling to bring its facts to life. The focus is collecting and transforming data into impactful stories that aid decision-making. This chapter will explore how UNESCO, a prominent governmental organisation, utilises Big Data to document, understand, safeguard, and promote intangible cultural heritage.

6.4 Big Data in CHT

6.4.1 Big Data storytelling in ICH – A practical example from UNESCO

This section presents the practical applications of Big Data in the storytelling of Intangible Cultural Heritage (ICH) tourism. We begin with an overview of the UNESCO Intangible Cultural Heritage Dataset, emphasising its role in Big Data analysis for ICH. The section covers various Big Data analysis types, preparation methods, and techniques like correlation analysis. A key focus is on storytelling through Big Data, highlighting how data narratives can enrich our understanding of ICH.

We spotlight the UNESCO initiative "Dive into Intangible Cultural Heritage", an exemplary model of Big Data storytelling. This project uses web semantics and graphical visualisation to engage with global ICH, featuring a semantic graph-based dataset linking ICH elements, NGOs, and scientific publications. The goal is to explain ICH and demonstrate the connections between cultural heritage elements with a dynamic, visually rich resource.

The [ICH Constellation](), accessible via the UNESCO website, is another focal point. It visually represents the diversity and interconnectivity of ICH elements, offering an interactive experience to explore themes like family rituals and handicrafts (Figure 6.2). This tool helps contextualise ICH about themes like sustainable development and potential threats, making it a valuable resource for cultural heritage and tourism stakeholders.

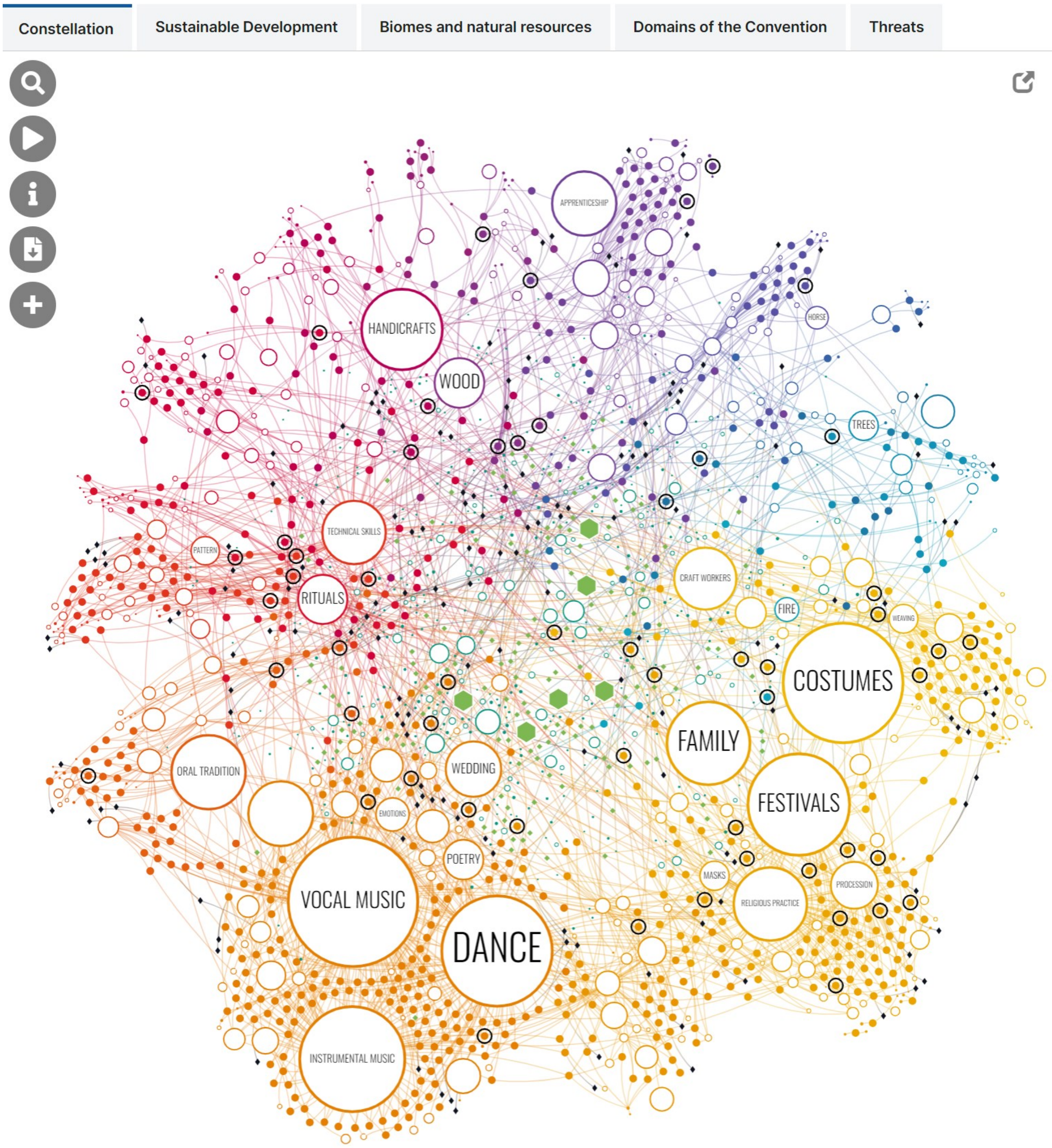

Figure 6.2: The UNESCO "Diving into Intangible Cultural Heritage" Constellation (Source: UNESCO Dive into Intangible Cultural Heritage website)

6.4.2 Big Data Storytelling in Tourism – Practical Example from UNTWO

The United Nations World Tourism Organisation (UNWTO) focuses on understanding global tourism for strategic planning and management. Its objectives include generating market knowledge, promoting sustainable tourism policies, enhancing tourism education, and using tourism as a development tool. The UNWTO collects extensive data on various aspects of tourism, which governments, organisations, and academics utilise. To make this Big Data accessible, it publishes annual reports summarising tourism data and offers the UNWTO Tourism Data Dashboard. This online tool allows users to quickly access and interpret tourism data, for example, inbound tourist statistics for specific countries and years, as demonstrated in Figure 2.3.

Figure 6.3: Screenshots of the UNWTO Tourism Dashboard and use (Source: UNWTO Tourism Dashboard Webpage)

So far, we've explored advanced examples of Big Data storytelling in Intangible Cultural Heritage (ICH) and tourism. The final sections will focus on more straightforward storytelling methods, using simpler ICH and tourism datasets analyses. This approach aims to make data storytelling more accessible and understandable.

6.4.3 Big Data Storytelling in ICH CHT – self-created example

To illustrate our analysis, we merged two datasets: the UNESCO Intangible Cultural Heritage (ICH) dataset (can be accessed for free by clicking here), detailing ICH in 128 countries over 13 years with 562 elements, and the UNWTO tourism dataset (access it by clicking here), recording tourist arrivals in 218 countries from 1995 to 2020. After aligning ICH data with corresponding countries, we removed incomplete entries, notably from 2020 due to COVID-19. The ICH was categorised into Music, Dance, Spiritual, Events, and Food for manageability. The final dataset, covering 103 countries from 2008 to 2019, provided a comprehensive view of ICH practices and their correlation with tourism.

6.4.3.1 Statistical approach

Through a statistical analysis (correlation), we could determine the correlation between the numbers of ICH per country and the changes in arriving tourists. This will help us see if there is a correlation between the increase in identified ICH elements and an increase in arriving tourists. In Figure 6.4, the correlation coefficients show a positive relationship between the number of ICH elements and the rise in tourist arrivals over the years.

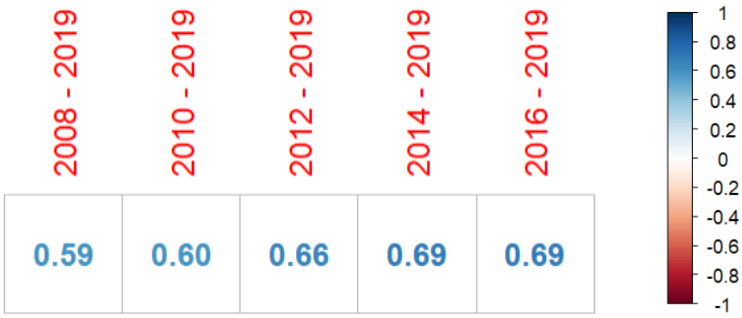

Figure 6.4: Correlation analysis (Changes in ICH elements vs. changes in tourist arrivals over time intervals) (Source: Zdanavičiūtė, Scholtz & De Ridder (2023) - YouTube)

Afterwards, we also examined the correlation between the increases in the registered different ICH element groupings and changes in tourist arrivals per year (see Figure 6.5).

Analysis in Figure 6.5 reveals the most significant correlations are between 'Events ICH' and tourist numbers, followed by 'Music ICH' in 2012-2019 and 2014-2019. Furthermore, a growing correlation is observed between 'Food ICH' and tourist figures from 2016 to 2019.

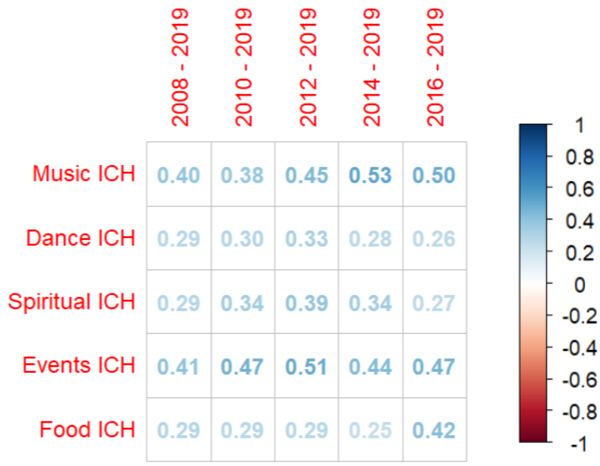

Figure 6.5: Correlation analysis (Changes in ICH element groupings vs changes in tourist arrivals over time intervals) (Source: Zdanavičiūtė, Scholtz & De Ridder (2023) - YouTube)

While these findings are apparent to those with a statistical background, the average reader may find understanding these analyses challenging without additional narrative context. This underscores the importance of storytelling in making statistical data accessible and engaging to a broader audience.

6.4.3.2 The effects of a lack of storytelling

By excluding even the most basic forms of storytelling, the results become more difficult to interpret. This is due to a few reasons:

- **Understanding takes more time:**
  - The purpose of the analysis or what is being portrayed is not immediately apparent.
  - Viewers' attention span may be short.
  - The importance/meaning of the content can be lost.
- **Statistical knowledge:**
  - Individuals with less statistical knowledge might have trouble understanding.
- **Increased interpretation and explanation**
  - More text and verbal communication are needed to create understanding.
  - Possibility of leaving out or missing crucial contextual information.
- **Collectively, these can lead to:**
  - Inaccurate conclusions.
  - Obvious patterns are not picked up on or misinterpreted.
  - Poor business decisions.

6.4.3.3 The application of basic storytelling

In this chapter, we have explored storytelling in data presentation, emphasising simplicity and visual impact. Complex data is made accessible and engaging by using images, icons, and varied sizes on a timeline. For example, visualising the correlation coefficient not as numbers but as dynamic images; a more significant number corresponds to a larger image. This approach is illustrated by the growth of Intangible Cultural Heritage (ICH) practices, represented by an expanding UNESCO Convention for the Safeguarding of Intangible Cultural Heritage logo and increasing tourist numbers depicted by multiplying person icons. Placed on a timeline, these visual cues facilitate intuitive understanding with minimal explanation.

Data visualisation (Figure 6.6) reveals a parallel increase in ICH practices and tourism since 2008, aligning with UNESCO's Representative List establishment. Annually, approximately 38 practices are added. Although this concurrent rise in heritage listings and tourist numbers doesn't imply causation, it indicates a potential relationship worth further investigation, perhaps exploring how ICH influences tourism growth.

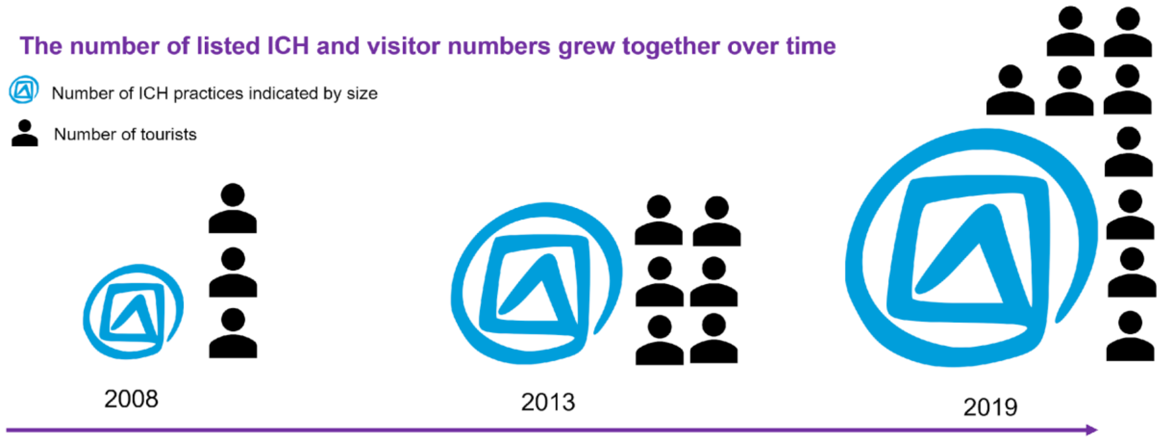

Figure 6.6: Similarities regarding an increase in visitor numbers and the number of registered ICH types (Source: Zdanavičiūtė, Scholtz & De Ridder (2023) - YouTube)

Delving into Figure 6.7, we observe that 'intangible cultural heritage' (ICH) spans a variety of forms like music, dance, spiritual rituals, cuisine, and events, categorised into five principal groups. Notably, events and music exhibit a pronounced correlation with tourism figures, with gastronomy also showing a notable link. This pattern indicates these heritage forms, especially events and music, as potentially significant tourist attractions. However, further research is necessary to determine causal[12] relationships and ensure that tourism development responsibly supports and sustains these cultural practices.

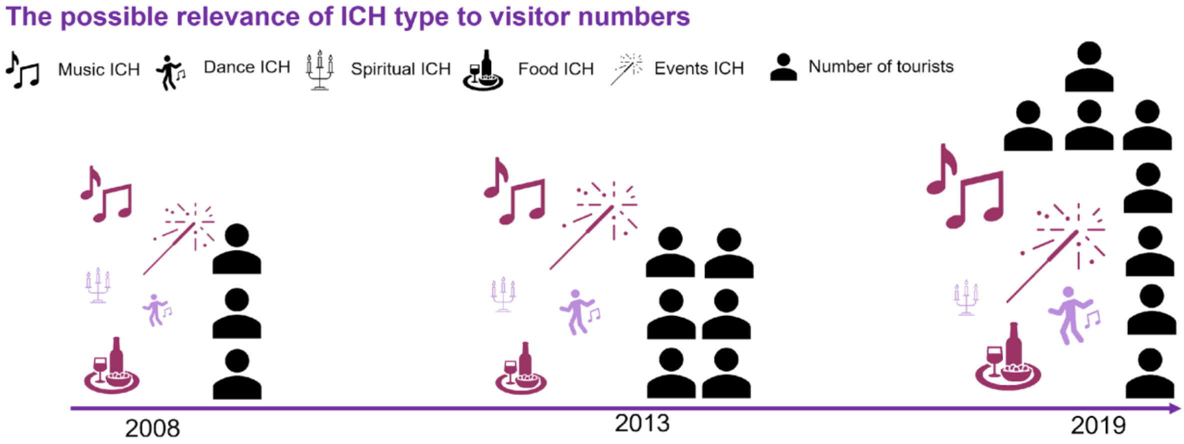

Figure 6.7: Similarities regarding an increase in visitor numbers and the number of registered ICH-type groupings over the years (Source: Zdanavičiūtė, Scholtz & De Ridder (2023) - YouTube)

6.5 Conclusion

In conclusion, this chapter highlights the transformative role of Big Data in enhancing cultural heritage tourism. By integrating vast datasets with storytelling, tourism and heritage professionals can uncover more profound insights into visitor behaviour and preferences, leading to more personalised and meaningful experiences that simultaneously foster responsible behaviour and sustainable practices and facilitate improved safeguarding. While data collection and analysis challenges persist, Big Data's potential in shaping future tourism strategies is immense. Embracing this technology not only aids in safeguarding intangible cultural heritage but also ensures its sustainable and dynamic promotion in an ever-evolving global tourism landscape. As the field continues to advance, the synergy between Big Data and storytelling in tourism is set to redefine how we understand and experience cultural heritage.

[12] A 'causal relationship' refers to a situation where one thing is proven to directly cause another. It means there's evidence that changes in one variable (the cause) directly bring about changes in another (the effect).

Chapter 7: The Future of Big Data Storytelling Techniques

Dr Luca Stornaiuolo
Co-founder Toretei
Milan, Italy

TO WATCH THIS CHAPTER VIA WEBINARS, CLICK HERE

7.1 EMERGING TECHNOLOGIES IN DATA VISUALISATION

Extending the research on emerging technologies in data visualisation involves exploring the latest advancements, tools, and platforms shaping the future of how we interact with and understand data. These technologies enhance the aesthetics of data presentation and make complex data sets more accessible and engaging for a broader audience. Here's a deeper look into this area's current state and potential future developments.

7.1.1 AUGMENTED REALITY (AR) AND VIRTUAL REALITY (VR) IN DATA VISUALISATION

AR and VR technologies transform data visualisation by creating immersive experiences. For instance, Microsoft's HoloLens and other AR glasses allow users to interact with data in 3D, making complex information more intuitive to understand. VR applications like those developed using Unity or Unreal Engine enable users to explore their data in a fully immersive environment. As AR and VR technologies become more accessible and affordable, we can expect these immersive data experiences to become more common in education, healthcare, and urban planning. This could revolutionise how we learn from and interact with data, providing a deeper understanding of complex systems like human anatomy, climate change models, and architectural designs.

7.1.2 ARTIFICIAL INTELLIGENCE (AI) AND MACHINE LEARNING (ML) IN DATA VISUALISATION

AI and ML are already being used to automate the analysis of large data sets and to identify patterns and insights that would be difficult for humans to find. Tools like Google's AutoML and IBM's Watson can analyse data and generate visualisations without requiring users to sift through the data manually.

Looking ahead, AI could offer personalised data visualisation experiences by understanding the user's context and preferences. For example, an AI system could automatically adjust the complexity and format of a visualisation based on the user's expertise level or specific interests. This personalisation could make data more accessible and engaging for a wider range of audiences.

7.1.3 Real-time Data Visualisation

Technologies such as web sockets and real-time databases are enabling the development of live data visualisation tools. These tools can display updated data in real-time, providing immediate insights into changing conditions. Examples include live dashboards for monitoring stock market trends, traffic conditions, and social media sentiment. As the Internet of Things (IoT) expands, real-time data visualisation will become increasingly important for monitoring and managing the vast amount of data generated by connected devices. This could lead to advancements in smart city technologies, real-time health monitoring systems, and dynamic supply chain management solutions.

7.1.4 Interactive Data Visualisation

Tools like Tableau, Power BI, and D3.js enable users to create interactive data visualisations that can be explored dynamically. These visualisations allow users to drill down into specific data points, adjust parameters, and view data from different angles. Future developments in interactive data visualisation could include more intuitive interfaces for exploring data, such as natural language processing (NLP) capabilities that allow users to ask questions about the data in plain language. Additionally, advances in AI could enable these tools to suggest new ways of looking at the data based on the user's interactions and interests.

7.1.5 Collaborative Data Visualisation

Cloud-based platforms like Google Data Studio and Microsoft Power BI make collaborating on data visualisation projects easier for teams. These platforms allow multiple users to work on the same visualisation simultaneously, share their work with stakeholders, and receive feedback in real-time. As collaboration becomes increasingly important in data-driven decision-making, we expect more sophisticated tools for visualisation of collaborative data. These might include features for version control, integrated communication tools, and enhanced security measures for sharing sensitive data.

By exploring these emerging technologies in data visualisation, we can anticipate a future where data is not only more accessible and understandable but also a more integral part of our daily decision-making processes. The evolution of these technologies promises to enhance our ability to convey complex narratives through data, making big data storytelling an even more powerful tool for communication, education, and social change.

7.2 Artificial Intelligence (AI) and Machine Learning

The role of Artificial Intelligence (AI) and Machine Learning (ML) in data visualisation and storytelling is rapidly evolving, offering unprecedented capabilities for analysing vast datasets, uncovering hidden patterns, and generating insightful narratives. This section delves deeper into how AI and ML are transforming the landscape of big data storytelling, focusing on current applications and speculating on future trends.

7.2.1 Automated Insight Generation

Automated insight generation tools like IBM Watson and Google Cloud AI leverage natural language processing (NLP) and machine learning algorithms to analyse data and extract meaningful insights automatically. These tools can identify trends, anomalies, and correlations within large datasets and present them in an understandable format. Narrative Science's Quill is an advanced natural language generation platform that transforms data into narrative reports, summarising complex information into easily digestible text. This technology is used across finance, marketing, and sports to generate reports, news stories, and personalised narratives from structured data.

7.2.2 Predictive Analytics and Forecasting

Predictive analytics use historical data to predict future events, trends, and behaviours. Machine learning models are trained on past data to forecast outcomes with a certain probability, aiding decision-makers across various fields, from finance to healthcare. In healthcare, predictive analytics can forecast disease outbreaks or patient readmissions, enabling proactive measures. Tools like Health Catalyst leverage ML algorithms to predict clinical outcomes and optimise patient care.

7.2.3 Enhancing Data Visualisation with ML

ML algorithms can enhance data visualisation by identifying the most relevant features of the data to be highlighted. This approach helps create more effective and focused visual representations, ensuring that users are not overwhelmed by the complexity of the data.

Google's Facets, an open-source visualisation tool, allows users to explore and analyse machine learning datasets interactively. It uses ML to help users understand and visualise data distribution across different dimensions, facilitating more informed analysis.

7.2.4 Interactive and Dynamic Visualisations

AI is enabling the creation of more interactive and dynamic visualisations that respond to user inputs or changes in data in real-time. These visualisations can adapt to provide personalised insights based on the user's preferences or queries. Tableau's "Ask Data" feature allows users to interact with their data using natural language queries. The tool uses NLP to understand the query and ML to generate the most relevant visualisations, making data exploration more intuitive.

7.2.5 Ethical Considerations and Bias Mitigation

As AI and ML become more integrated into data storytelling, ethical considerations, particularly regarding bias and fairness, have come to the forefront. Developers and researchers are working on methods to detect and mitigate biases in AI algorithms to ensure fair and accurate representations of data. IBM's AI Fairness 360 is an open-source toolkit designed to help detect and mitigate bias in ML models. It provides a comprehensive set of metrics and algorithms that can be applied to improve fairness in various stages of the AI lifecycle.

7.2.6 Future Trends

- **Explainable AI (XAI)**

As AI systems become more complex, a growing demand for explainable AI allows users to understand and trust how decisions are made. XAI aims to make AI decision-making processes transparent and interpretable, enhancing the credibility of AI-generated insights in data storytelling.

- **Generative Models for Data Synthesis**

Generative models like Generative Adversarial Networks (GANs) could be used to synthesise data for storytelling, creating realistic but artificial datasets that can help illustrate potential future scenarios or highlight specific insights without using sensitive real-world data.

- **Augmented Creativity**

AI could assist in the creative aspects of storytelling, suggesting novel ways to visualise data or generating initial drafts of narratives that analysts and storytellers can refine and enhance.

Integrating AI and ML in big data storytelling is not just about making sense of large datasets but about unlocking the narrative potential of data in previously unimaginable ways. As these technologies continue to evolve, they will undoubtedly redefine the art and science of storytelling, making it more interactive, insightful, and inclusive.

7.3 Ethical Considerations in Data Storytelling

The integration of big data and storytelling, particularly when enhanced by technologies like AI and ML, raises significant ethical considerations and data privacy concerns. This section expands on these issues, exploring the balance between leveraging data for insightful storytelling and respecting individual privacy and data rights. It also delves into the ethical challenges of bias and misinformation in data narratives.

- **Transparency and Consent**

In the era of big data, obtaining explicit consent from individuals whose data is being collected, analysed, and narrated has become a cornerstone of ethical data use. Transparency about what data is collected, how it is used, and who has access to it is crucial. The European Union's General Data Protection Regulation (GDPR) sets a precedent for transparency and consent in data handling, offering a framework that many data storytellers outside the EU also choose to follow to uphold high ethical standards.

- **Avoiding Misinterpretation**

Ethical data storytelling involves presenting data in a way that is truthful and avoids misleading interpretations. This includes careful selection of visualisation techniques and narrative context to ensure that the story told by the data is accurate and not biased towards a particular conclusion. Misleading visualisations can create panic or complacency in the context of reporting on public health data. Ethical storytellers must choose graphs and scales that accurately represent the situation, such as accurately contextualising the numbers in a pandemic to avoid either underestimating or exaggerating the severity.

7.3.1 Data Privacy Concerns

- **Anonymisation and De-identification**

Protecting individual privacy involves anonymising datasets so that personal information cannot be traced back to individuals without their consent. This is particularly challenging in the age of big data, where multiple datasets can be combined to re-identify individuals. The release of anonymised taxi trip data by the New York City Taxi and Limousine Commission, which was later re-identified to expose individual drivers' earnings and habits, underscores the challenges of ensuring privacy in publicly shared data.

- **Secure Data Handling Practices**

Implementing secure data storage and transfer methods to protect against unauthorised access is a critical aspect of respecting data privacy. This includes encryption, secure authentication mechanisms, and regular audits of data access logs. The adoption of blockchain technology for secure, decentralised data storage in projects requiring high data integrity and privacy levels, such as medical records in healthcare research, illustrates advanced approaches to secure data handling.

7.3.2 Addressing Bias and Ensuring Fairness

- **Bias Detection and Mitigation**

AI and ML models can inadvertently perpetuate or even amplify biases present in the training data. Ethical data storytelling requires active efforts to detect and mitigate these biases to ensure that narratives do not reinforce stereotypes or injustices. AI ethics researchers are developing tools and methodologies, such as IBM's AI Fairness 360, to identify biases in machine learning models. These tools can help data storytellers ensure that their narratives are grounded in fair and unbiased data representations.

- **Diversity and Inclusion**

Including diverse perspectives in the data storytelling process can help identify potential biases and ethical concerns that might not be evident to a more homogenous team. This involves not only diversity in terms of demographics but also in expertise and thought. Projects like Google's "Project Euphonia" demonstrate the importance of including diverse perspectives by using AI to improve speech recognition systems for people with speech impairments, showing how considering varied user needs leads to more inclusive technology solutions.

7.3.3 Future Directions

- **Ethical Frameworks and Guidelines**

Developing and adhering to ethical frameworks and guidelines for data storytelling is becoming increasingly important. These frameworks can help organisations navigate the complex landscape of ethical considerations and data privacy concerns. The Data Ethics

Framework by the UK Government provides guidelines for ethical data use in public sector projects, offering a model that can be adapted for data storytelling initiatives.

- **Public Discourse and Education**

Engaging in public discourse about the ethical implications of data storytelling and educating data scientists, storytellers, and the public about these issues are crucial steps towards a more ethical data culture. Initiatives like the Markkula Centre for Applied Ethics at Santa Clara University's discussions and publications on data ethics are vital in raising awareness and fostering ethical practices in data use and storytelling. By addressing these ethical considerations and data privacy concerns, data storytellers can not only navigate the challenges posed by big data but also harness its potential to tell stories that are both impactful and respectful of individual rights and societal norms.

7.4 Interdisciplinary approaches to Data Science

Interdisciplinary approaches to data storytelling leverage the strengths and perspectives of various fields to enhance the way narratives are constructed and presented from complex datasets. This methodological fusion between data science, visual arts, journalism, psychology, and other disciplines enriches storytelling, making it more accessible, engaging, and impactful. Here, we explore how these interdisciplinary collaborations are shaping the future of data storytelling.

7.4.1 Journalism and Data Science

Data journalism represents a blend of investigative reporting and data science, where journalists use data analysis and visualisation tools to uncover and tell stories hidden within datasets. The Guardian's Datablog and The New York Times' Upshot are pioneering examples of data journalism. They utilise data visualisations and analyses to provide deeper insights into current events, ranging from election results and economic trends to public health crises.

7.4.2 Visual Arts and Data Visualisation

Artists and graphic designers collaborate with data scientists to create visual data narratives that are informative and aesthetically compelling. This collaboration aims to make data more relatable and understandable to a broader audience by employing principles of design and storytelling. Giorgia Lupi's data-driven art projects, such as "Dear Data" and collaborations with fashion brands and Other Stories, showcase how artistic sensibilities can bring data to life in unique and engaging ways.

7.4.3 Psychology and User Experience (UX) Design

Understanding how people perceive and interact with data is crucial for effective storytelling. Psychology and UX design play significant roles in creating data visualisations that are intuitive and emotionally resonant, ensuring that the narratives are not only seen but felt and understood. The interactive "Out of Sight, Out of Mind" visualisation of drone strikes by Pitch Interactive uses psychological principles to evoke empathy and understanding, making abstract numbers more tangible and emotionally impactful.

7.4.4 Computer Science and Interactive Technologies

Advances in computer science, particularly in areas like natural language processing (NLP), machine learning (ML), and interactive web technologies, have expanded the toolkit available for data storytellers. These technologies enable dynamic and personalised storytelling experiences. Google's "Talk to Books" project employs NLP to allow users to interact with books through conversational queries, demonstrating how interactive technologies can create new forms of engagement with information.

7.4.5 Ethnography and Cultural Studies

Incorporating ethnographic research and cultural studies into data storytelling helps contextualise data within people's lived experiences. This approach ensures that stories are grounded in the social and cultural dimensions that data alone cannot capture. The COVID-19 Mobility Data Network, a collaboration between epidemiologists and social scientists, uses mobility data to understand and combat the spread of COVID-19 while considering the socio-economic implications of the pandemic.

7.4.6 Future Directions

- **Collaborative Platforms and Tools**

The development of collaborative platforms and tools that facilitate interdisciplinary work is crucial for the future of data storytelling. These platforms must support diverse datasets, visualisation techniques, and real-time collaboration across different fields.

- **Education and Training Programs**

As the demand for interdisciplinary approaches grows, educational institutions are beginning to offer programs that cross-train students in data science, journalism, design, and other relevant fields. This trend is likely to continue, producing a new generation of storytellers equipped to navigate the complexities of big data.

- **Ethical and Inclusive Narratives**

Interdisciplinary approaches also bring diverse perspectives to the forefront, encouraging more ethical and inclusive narratives. Data storytellers can ensure their narratives reflect a broader range of experiences and viewpoints by integrating sociology, anthropology, and ethics insights. Interdisciplinary approaches to data storytelling enrich the narrative possibilities and democratise data, making it accessible and engaging for all. By breaking down silos between disciplines, storytellers can unlock innovative ways to visualise and communicate complex information, fostering a more informed and empathetic society.

7.5 Personalisation and Audience Engagement

Personalisation and audience engagement are critical components in the evolution of data storytelling, transforming how narratives are tailored and delivered to meet the unique preferences and needs of diverse audiences. This approach leverages data analytics, machine learning, and user experience design to create stories that resonate personally, enhancing comprehension, retention, and impact. Here, we delve deeper into the strategies and technologies driving this trend and explore examples of personalised data storytelling in action.

7.5.1 Strategies for Personalised Storytelling

- **Audience Segmentation**

Utilising data analytics to segment the audience based on demographics, interests, behaviour, or other relevant criteria enables storytellers to tailor narratives that align with each segment's specific characteristics and needs.

- **Adaptive Content**

Implementing machine learning algorithms to dynamically adjust a story's content, format, or presentation style based on real-time user interactions or feedback. This ensures the storytelling experience evolves to match the user’s preferences or learning pace.

- **Interactive Elements**

Incorporating interactive elements into data visualisations allows users to explore data points and narratives that interest them the most. This can include clickable maps, sliders to adjust time frames, or filters to examine specific data subsets.

7.5.2 Technologies Enabling Personalised Storytelling

➢ **Natural Language Processing (NLP)**

NLP technologies facilitate more natural interactions with data, allowing users to ask questions or express interests in conversational language. This can guide the storytelling system to generate personalised narratives based on user queries.

➢ **Artificial Intelligence (AI) and Machine Learning (ML)**

AI and ML are at the heart of personalisation, analysing user data to predict preferences and automatically generate customised content. These technologies can identify patterns in user behaviour to suggest relevant data stories or visualisations.

➢ **Web Analytics and Tracking Tools**

Tools like Google Analytics enable data storytellers to understand how users interact with their content, providing insights into which stories engage users the most. This data can inform the development of future narratives that better align with audience interests.

7.5.3 Examples of Personalised Data Storytelling

➢ **Spotify Wrapped**

An annual campaign by Spotify that provides users with personalised summaries of their listening habits over the year. Using data analytics and visualisation, Spotify creates engaging, shareable stories about each user's favourite artists, songs, and genres, demonstrating the power of personalised storytelling to connect on an individual level.

➢ **The New York Times Interactive Stories**

The New York Times has developed several interactive data stories that allow readers to input their information or make selections to see how broader data trends apply to them personally. Examples include interactive quizzes on dialects and personal finance calculators that offer insights based on user-provided data.

➢ **Strava Fitness Reports**

Strava, a social network for athletes, provides personalised yearly summaries highlighting an individual's athletic achievements, trends in activity levels, and comparisons with past performance. These reports use data visualisation to create a customised narrative of the user's fitness journey.

7.5.4 Future Directions

➢ **Ethical Considerations in Personalisation**

As personalisation in data storytelling advances, ethical concerns about privacy, data security, and consent become increasingly important. Future developments must balance the desire for personalised experiences with the imperative to protect user data and ensure transparency in how personal data is used.

➢ **Cross-Platform Personalisation**

Looking ahead, personalised data storytelling will likely extend across multiple platforms and devices, offering a seamless and integrated user experience. This will require data integration and cross-platform analytics advancements to ensure consistent and relevant storytelling across different user touchpoints.

➢ **Enhanced User Models**

Future personalisation efforts will benefit from more sophisticated user models incorporating a more comprehensive range of data, including emotional responses, cognitive styles, and social context. This could lead to even more nuanced and responsive storytelling experiences that adapt to what users want to see and how they feel and learn best. Personalisation and audience engagement in data storytelling represent a significant shift towards more user-centric narratives. By leveraging advanced analytics, AI, and interactive technologies, storytellers can create experiences that inform and connect with audiences on a deeply personal level, making data more relevant, engaging, and impactful.

7.6 Future Challenges

The future of big data storytelling is rife with both challenges and opportunities. Storytellers must navigate a complex landscape as the field continues to evolve, driven by technological advancements and changing societal needs. This section explores key challenges and opportunities that will shape the future of data storytelling, offering insights into how storytellers can adapt and thrive.

- **Data Overload and Complexity**

One of the primary challenges in big data storytelling is the sheer volume and complexity of data available. As data grows exponentially, finding meaningful stories amidst the noise becomes increasingly tricky. Storytellers must develop advanced analytical skills and leverage sophisticated tools to distil complex datasets into coherent, engaging narratives.

- **Maintaining Accuracy and Trust**

In an era of misinformation and "fake news," maintaining the accuracy and trustworthiness of data stories is paramount. As data manipulation becomes more sophisticated, storytellers are responsible for ensuring their narratives are rooted in verifiable facts and presented in a way that fosters trust.

- **Ethical Use of Data**

With great power comes great responsibility. The ethical implications of how data is collected, analysed, and presented are becoming more pressing. Storytellers must navigate privacy concerns, consent, and the potential for bias in their narratives, ensuring they respect individual rights and societal norms.

- **Bridging the Skills Gap**

The interdisciplinary nature of big data storytelling requires a blend of skills, including data analysis, narrative construction, visual design, and technological proficiency. Bridging the gap between these diverse skill sets presents a significant challenge for individuals and organisations.

7.6.1 Future Opportunities

- **Advancements in AI and ML**

The continued development of AI and ML technologies offers exciting opportunities for automating parts of the data analysis and storytelling process. These technologies can help identify patterns, generate insights, and even craft preliminary narratives, allowing storytellers to focus on the creative aspects of their work.

- **Interactive and Immersive Storytelling**

Emerging technologies like AR, VR, and interactive web platforms open new avenues for immersive storytelling. These technologies can create more engaging and personalised narrative experiences, allowing audiences to explore data stories more effectively and meaningfully.

- **Democratisation of Data Storytelling Tools**

The development of user-friendly data visualisation and storytelling tools has the potential to democratise data storytelling, making it accessible to a broader audience. This democratisation can empower people to tell their own data-driven stories, fostering a more informed and engaged society.

- **Cross-disciplinary Collaborations**

The future of data storytelling lies in cross-disciplinary collaborations that bring together expertise from data science, journalism, design, psychology, and other fields. These collaborations can enrich data narratives, making them more nuanced, impactful, and relevant to diverse audiences.

- **Global and Cultural Narratives**

As data storytelling becomes more global, there is an opportunity to tell cross-cultural and international stories that highlight shared experiences and challenges. This global perspective can foster empathy, understanding, and collaboration across borders.

7.6.2 Navigating the Future

To navigate these challenges and capitalise on the opportunities, data storytellers must stay agile, continuously updating their skills and adapting to new technologies and societal changes. They will also need to foster a culture of ethical storytelling, prioritising accuracy, transparency, and respect for privacy. Educational institutions and organisations can play a critical role in preparing the next generation of data storytellers by offering interdisciplinary training and fostering an environment of continuous learning and collaboration.

The future of big data storytelling is not without its hurdles, but it also promises more engaging, insightful, and impactful narratives. By embracing the challenges and opportunities, storytellers can help shape a future where data-driven narratives inform, inspire, and drive positive change.

Reference lists

Chapter 1 references

Augmenting Data Warehouses with Big Data. (April 2015). Information Systems Management, 32(3):150428173206006. DOI: 10.1080/10580530.2015.1044338.

Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a Literature-Driven Definition of Big Data in Healthcare. PMID: 26137488 PMCID: PMC4468280 DOI: 10.1155/2015/639021.

Beyer, M.A., & Laney, D. (2012). The Importance of “Big Data”: A Definition. Gartner. Retrieved from https://www.gartner.com/doc/2057415.

Boldosova, V., & Luoto, S. (2002). Storytelling, Business Analytics and Big Data Interpretation: Literature Review and Theoretical Propositions. Management Research Review, 43(2), 204-222. doi:10.1108/MRR-03-2019-0106.

Bresciani, S., & Eppler, M. J. (2009). The Benefits of Synchronous Collaborative Information Visualization: Evidence from an Experimental Evaluation. IEEE Transactions on Visualization and Computer Graphics, 15(6).

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. Pages 662-679. https://doi.org/10.1080/1369118X.2012.678878.

Charon, R. (2008). Narrative Medicine: Honoring the Stories of Illness. Oxford University Press.

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. Information Sciences, 275, 314-347. DOI: 10.1016/j.ins.2014.01.015.

Chen, H., Chiang, R.H., & Storey, V.C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36, 1165-1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19, 171-209. https://doi.org/10.1007/s11036-013-0489-0.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A Formal Definition of Big Data Based on Its Essential Features. Library Review, 65(3), 122-135. DOI: 10.1108/LR-06-2015-0061.

Dumbill, E. (2013). Making Sense of Big Data.

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is Your Definition of Big Data? Researchers' Understanding, Practical Definitions and Socio-Cultural Evolutions.

Few, S. (2009). Now You See It: Simple Visualization Techniques for Quantitative Analysis. Analytics Press.

Gadinger, F., Mert, A., & Ochoa, C. S. (Eds.). (2016). Political Storytelling: From Fact to Fiction. Duisburg, Basis-Druck.

Green, M. C. (2004). Storytelling in Teaching. APS Observer, 17(4), 37-39.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (n.d.). The Rise of “Big Data” on Cloud Computing: Review and Open Research Issues.

Hollihan, T. A., Baaske, K. T., & Riley, P. (1987). Debaters as Storytellers: The Narrative Perspective in Academic Debate. The Journal of the American Forensic Association, 23(4), 184-193. DOI:10.1080/00028533.1987.11951344.

Jonas, C. (2002). Playing the Story: Computer Games as a Narrative Genre. Human IT, 6(3), 7-53.

Kirk, A. (2016). Data Visualisation: A Handbook for Data Driven Design. SAGE Publications.

Kugelmann, R. (2001). Introducing Narrative Psychology: Self, Trauma and the Construction of Meaning. Journal of Health Psychology, 6(5), 604-606. doi:10.1177/135910530100600511.

Limaj, B. (2000). Big Data Systems: A Renewed Definition. Early 2000s.

Maddalena, F., Eva, D. C., Christophe, O. S., & Bernice, S. E. (2020). What is Your Definition of Big Data? Researchers' Understanding, Practical Definitions and Socio-Cultural Evolutions.

Matrix, S. (2014). The Netflix Effect: Teens, Binge Watching, and On-Demand Digital Media Trends. Jeunesse: Young People, Texts, Cultures, 6(1), 119-138. doi:10.1353/jeu.2014.0002.

Nelson, D. L., Reed, V. S., & Walling, J. R. (1976). Pictorial Superiority Effect. Journal of Experimental Psychology: Human Learning and Memory, 2(5), 523–528. https://doi.org/10.1037/0278-7393.2.5.523.

Nicolosi, G., Korthals, M. (2008). Narrative Strategies in Food Advertising. In Coff, C., Barling, D., Korthals, M., Nielsen, T. (Eds.), Ethical Traceability and Communicating Food. Berlin, Springer.

Presser, L., & Sandberg, S. (2015). Narrative Criminology: Understanding Stories of Crime. New York: NYU Press.

Pulizzi, J. (2012). The Rise of Storytelling as the New Marketing. Pub Res Q, 28, 116–123. doi:10.1007/s12109-012-9264-5.

Richter, A., Sieber, A., Siebert, J., Miczajka-Rußmann, V. L., Zabel, J., Ziegler, D., Hecker, S., & Frigerio, D. (2019). Storytelling for Narrative Approaches in Citizen Science: Towards a Generalized Model. JCOM, 18(06), 1-24. doi:10.22323/2.18060202.

Shiller, R. J. (2019). Narrative Stories: How Stories Go Viral & Drive Major Economic Events. Princeton University Press.

Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics Press.

Undefined By Data: A Survey of Big Data Definitions. (n.d.). Jonathan Stuart Ward, Adam Barker. Retrieved from https://doi.org/10.48550/arXiv.1309.5821.

Williams, S. (2020). Data Action: Using Data for Public Good. The MIT Press.

Zak, P. J. (2014). Why Your Brain Loves Good Storytelling. Harvard Business Review.

Chapter 2 references

Alon Halevy, Peter Norvig, and Fernando Pereira, ”The Unreasonable Effectiveness of Data”, (2009).

Eugene Wigner, ”The Unreasonable Effectiveness of Mathematics in the Natural Sciences” by Eugene Wigner (1960).

Martin Gardner, ”The Surprising Birthday Problem”, Scientific American (1976) [4] Abhinav Gupta, ”Revisiting the Unreasonable Effectiveness of Data”, Google Blog, 2017

Chapter 3 references

Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter - Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (Accessed: January 8, 2024).

Searle, J.R. (1980). “Minds, Brains, and Programs”, The Behavioral and Brain Sciences, 3: pp. 417-424.

Chapter 4 references

Lengler, R., & Eppler, M. J. (2007, January). Towards a periodic table of visualization methods for management. In IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA (Vol. 7, No. 8, p. 31). URL http://www.visual-literacy.org/periodic_table/periodic_table.pdf.

Wikipedia. Histogram — Wikipedia, the free encyclopedia. URL https://en.wikipedia.org/wiki/Histogram#/media/File:Cumulative_vs_normal_histogram.svg

Wikipedia. Box plot — Wikipedia is a free encyclopedia. URL http://en.wikipedia.org/w/index.php?title=Box_plot&oldid=514024080

G. Dzemyda, O. Kurasova, and J. Žilinskas. *Daugiamaciu duomenu vizualizavimo metodai*. Vilnius: Matematikos ir informatikos institutas, 2008.

https://datavizcatalogue.com/methods/scatterplot.html

https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/introduction-to-graph-theory/graph-theory-adjacency-matrices/

Wikipedia. Mosaic plot — wikipedia, the free encyclopedia. URL https://en.wikipedia.org/wiki/Mosaic_plot#/media/File:Mosaic-big.png

https://www.fusioncharts.com/dev/chart-guide/standard-charts/sankey-diagram

Wikipedia. Parallel coordinates — Wikipedia is a free encyclopedia. URL https://en.wikipedia.org/wiki/Parallel_coordinates#/media/File:ParCorFisherIris.png

Chen, C. H., Härdle, W. K., & Unwin, A. (Eds.). (2007). Handbook of data visualisation. Springer Science & Business Media.

Mackute-Varoneckiene, A., Zilinskas, A., & Varoneckas, A. (2009, June). Multidimensional scaling: a multi-objective optimisation approach. In Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing (pp. 1-6).

Wilke, C. O. (2019). Fundamentals of data visualisation: a primer on making informative and compelling figures. O'Reilly Media.

CHAPTER 5 REFERENCES

Bart de Langhe and Stefano Puntoni, Leading With Decision-Driven Data Analytics, MIT Slogan Management Review, December 07, 2020

CHAPTER 6 REFERENCES

Boldosova, V. (2019). Deliberate storytelling in Big Data analytics adoption. *Information Systems Journal*, *29*(6), 1126-1152.

Buhalis, D., & Sinarta, Y. (2019). Real-time co-creation and nowness service: lessons from tourism and hospitality. *Journal of Travel & Tourism Marketing, 36*(5), 563-582.

Doesselaere, E. (2020). *Immaterieel erfgoed borgen*. FARO. https://faro.be/kennis/immaterieel-erfgoed-borgen

Epamanywhere. (2022). Big Data Analytics in the Tourism and Travel Industry. Retrieved from: https://anywhere.epam.com/business/big-data-in-tourism-and-travel

Erfgoedcellen.be. (n.d.). Wat is erfgoed? Retrieved from: https://www.erfgoedcellen.be/wat-is-erfgoed/

Grepsr. (2021, August 11). Significance of Big Data in the Tourism Industry. Retrieved from https://www.grepsr.com/blog/big-data-tourism-industry/

Hall, C.M., & Lew, A.A. (2009). Understanding and managing tourism impacts: an integrated approach. (*1st ed.*). London, UK: Routledge

Kirshenblatt-Gimblett, B. (1998). *Destination Culture: Tourism Museums, and Heritage.* University of California Press.

Mariani, M. M., Baggio, R., Fuchs, M., & Höpken, W. (2018). Business intelligence and Big Data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management, 30*(12), 3514-3554.

Mariani, M., & Baggio, R. (2022). "Big Data and analytics in hospitality and tourism: a systematic literature review", International Journal of Contemporary Hospitality Management, *34*(1), 231-278.

Masoud, H., Mortazavi, M., & Farsani, N. (2019). A study on tourists' tendency towards intangible cultural heritage as an attraction (case study: Isfahan, Iran). City, Culture and Society, *17*, 54-60. https://doi.org/10.1016/j.ccs.2018.11.001

Mountasser, I., Ouhbi, B., Frikh, B., & Hdioud, F. (2020). Big Data research in the tourism industry: requirements and challenges. International Journal of Mobile Computing and Multimedia Communications (IJMCMC), 11(4), 26-41.

Octoparse. 2021. What is Big Data in Tourism? (Use cases explained). https://www.octoparse.com/blog/big-data-in-tourism#

Optimizely. 2022. The complete guide to Big Data analysis in travel. https://www.optimizely.com/insights/blog/the-complete-guide-to-big-data-analysis-in-travel/#:~:text=Travel%2Drelated%20companies%20that%20use,companies%20without%20these%20key%20insights

Pangaea X Blog. (2022, February 7). How Is Data Analytics Used in Tourism? Retrieved from https://www.pangaeax.com/blog/how-is-data-analytics-used-in-tourism/

Petronela, T. (2016). The importance of the intangible cultural heritage in the economy. *Procedia Economics and Finance, 39*, 731-736. https://doi.org/10.1016/S2212-5671(16)30271-4

UNESCO (2009). *Identifying and Inventorying Intangible Cultural Heritage.* Retrieved from https://ich.unesco.org/doc/src/01856-EN.pdf

UNESCO (2011). 2003 Convention for the Safeguarding of the Intangible Cultural Heritage. *Sixth session of the Intergovernmental Committee for Safeguarding of the Intangible Cultural Heritage.* 22-29 November 2011, Bali, Indonesia. Retrieved from: https://ich.unesco.org/doc/src/15164-EN.pdf

UNESCO. (2022). Basic texts of the 2003 Convention of Safeguarding the Intangible Cultural Heritage. Retrieved from https://ich.unesco.org/doc/src/2003_Convention_Basic_Texts-_2022_version-EN_.pdf

United Nations Environment Programme (UNEP). (2021). Travel and tourism industry chart new, greener course at COP26. Retrieved from https://www.unep.org/news-and-stories/story/travel-and-tourism-industry-chart-new-greener-course-cop-26

United Nations World Travel & Tourism Council (WTTC). (2023). Economic Impact Research. Retrieved from https://wttc.org/research/economic-impact

Volo, S. (2020). Tourism statistics, indicators and Big Data: a perspective article. Tourism Review, 75(1), 304-309.

Yallop, A., & Seraphin, H. (2020). Big Data and analytics in tourism and hospitality: opportunities and risks. Journal of Tourism Futures, 6(3), 257-262.

Chapter 7 references

Brynjolfsson, E., & McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W.W. Norton & Company. Retrieved from https://edisciplinas.usp.br/pluginfile.php/4312922/mod_resource/content/2/Erik%20-%20The%20Second%20Machine%20Age.pdf

Cairo, A. (2012). The Functional Art: An Introduction to Information Graphics and Visualization. New Riders. Retrieved from http://www.thefunctionalart.com/

Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books.

Duarte, N. (2019). DataStory: Explain Data and Inspire Action Through Story. Ideapress Publishing.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Kirk, A. (2019). Data Visualisation: A Handbook for Data Driven Design (2nd ed.). SAGE Publications Ltd.

Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE Publications Ltd.

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. John Murray.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

Peck, E. (2018). Personalized data storytelling with Tableau. Tableau Blog. Retrieved from https://www.tableau.com/en-gb/trial/visualize-your-data?d=7013y000002aJUqAAM&utm_campaign=Prospecting-CORE-ALL-ALL-ALL-ALL&utm_medium=Paid+Search&utm_source=Google+Search&utm_campaign_id=2017049&utm_language=EN&utm_country=BENX&adgroup=&adused=STAT&creative=&gclid=Cj0KCQiAoKeuBhCoARIsAB4WxtcLnkt0sAfH-bjKvDEDFO1_5hpFx3_URiE6-

UvroQU0zZVQgs9gjTQaAr_3EALw_wcB&gclsrc=aw.ds Richards, N. M., & King, J. H. (2014). Big data ethics. Wake Forest Law Review, 49(2), 393-432.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 1139-1148.