media numeric

Co-funded by the Erasmus+ Programme of the European Union

BY JACQUELINE PITSCH
& DANIEL SORABJI

MEDIANUMERIC

THE WORLD OF DATA-DRIVEN STORYTELLING IN JOURNALISM AND FACT-CHECKING

WHY IT IS IMPORTANT TO LEARN HOW TO WORK WITH LARGE DATASETS IN DIGITAL VERIFICATION AND WHAT BENEFITS THIS CAN BRING

CENTRUM CYFROWE · Storytek_ · SWPS University · SOUND&VISION · AFP · ina · nh · EU SCREEN

Colophon

About this report

Digital Storytelling In The Media Landscape is published by MediaNumeric, an initiative of the Netherlands Institute for Sound & Vision, InHolland University, L'Institut National de l'Audiovisuel, Agence France-Presse, SWPS University, Storytek, EUScreen and Centrum Cyfowe.

The ***Netherlands Institute for Sound & Vision*** is the institute for media culture; an inspiring, creative and accessible place for private individuals and professionals, focusing on current developments concerning people, media and society from a media-historical perspective.

InHolland University is a large university of applied sciences located in eight main cities in the Netherlands, offering practice-oriented education and research opportunities. They educate independent, critical professionals to make a meaningful contribution to the inclusive world of tomorrow, with a focus on fostering a sustainable living environment and a resilient society.

L'Institut National de l'Audiovisuel is a repository of all French radio and television audiovisual archives. Since 1974, Ina has been responsible for preserving, promoting and transmitting France's audiovisual heritage. Ina is also an international training and research centre for digital media and content.

Agence France-Presse (AFP) is one of the world's three major news agencies. Its mission is to provide rapid, comprehensive, impartial and verified coverage of the news and issues that shape people's daily lives.

SWPS University is a private non-profit university in Poland established in 1996, exploring the human mind and applying this expertise to address practical challenges in society, focusing on new technologies and dynamic social change.

Storytek Innovation & Venture Studio is a creative meditech and storytelling accelerator in Northern Europe, offering innovative business models and format options, as well as concrete needs for professional organisations in the European media ecosystem.

EUscreen is a network of European broadcasters and audiovisual archives, media scholars, and technical experts, facilitating access to and engagement with archival audiovisual content through their independent online portal.

Centrum Cyfrowe is a Polish think-and-do-tank supporting openness and engagement in the digital world by changing the way people learn, participate in culture, use the internet and exercise their rights as internet users.

About the authors

Jacqueline Pietsch is a Journalist and Technical Project Manager of Digital Investigations at AFP. She oversees technical developments for the team. Since joining AFP in 1996, she has worked in a myriad of roles, disciplines and countries, including head of AFP's English-language TV production, correspondent in the Netherlands and video journalist responsible for Southeast Asia.

Daniel Sorabji is Deputy Global Online Planning Coordinator at the London bureau of AFP. He is a creative media professional with a wide range of experience across various media-related disciplines including photography, photojournalism, editing, communications, planning, training and research.

Reviewers

Natalia Berger, Joke Hermes, Clément Malherbe, Sten-Kristian Saluveer, Rachel Somers Miles.

Cover Image

Designed by Rebecca Haselhoff, 2023.

Co-funded by the Erasmus+ Programme of the European Union

The MediaNumeric project has been co-funded by the European Commission under grant agreement No. 621610-EPP-1-2020-1-NL-EPPKA2-KA. This web site reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Table of Contents

1. Executive Summary

Digitisation has changed all aspects of the news media landscape, from the way content is created, to how it is distributed and interacted with. Three phenomena shape the face and fate of news media in Europe: decreasing trust and information disorder, digitisation and changing user behaviour, and dominance of global technology and Artificial Intelligence (AI). They rock the foundations of the journalistic profession.

The MediaNumeric programme provides students and young professionals in media and communication studies with the theoretical know-how and skills necessary to enable them to take on the opportunities of data-driven journalism and media production. It also highlights the potential of using large multimedia databases for data-driven innovations.

This deliverable describes the activities carried out during the three years of the MediaNumeric project within Work Package 2 (WP2) by the different partners and describes the results achieved by this work package. This deliverable D2.4 State of the Art Update, is in fact a further update of D2.2 Updated State of the Art of the Data-driven Journalism, taking the initial report of D2.2 updating the content and adding further chapters focusing on additional and new insights from this rapidly changing field.

WP2 D2.4 consists of an Updated State of the Art report that provides an overview of the world of data-driven storytelling in journalism and of digital verification of misinformation and disinformation, more commonly known as fact-checking. It also examines the impact of Artificial Intelligence (AI) on the news media and explores the role that Large Language Models could play in the organisation of databases. It also provides information for higher education institutions on best practices in how to teach these techniques.

The report examines why it is important that more people learn how to work with large datasets and in digital verification, who should learn these skills and what benefits this can bring to media organisations and indeed societies as a whole. It looks at the tools and best practices in training in this comparatively recent medium. It also examines why national audiovisual archives should embrace the role that they could play in providing rich datasets for research and insight into the past and present.

Among the key findings of this report are:

- Data-driven journalism brings high-value, high impact stories into the newsroom. These stories can attract a loyal readership and effect real-world change. Examples include the Roman Catholic child sex scandal uncovered by the *Boston Globe* and tax evasion as revealed by the ICIJ.

- Though misinformation and disinformation have existed since the dawn of time, the social media era is enabling information (and misinformation) to spread at split-second speeds, creating real-world harm as people turn to unqualified sources to seek answers to their questions.
- A fear of maths and spreadsheets and a perception that it is complicated are the biggest handicaps to more journalists adopting the skills needed to work with data.
- The most important tool in both the area of data-driven storytelling, artificial intelligence tools and debunking misinformation is knowing what questions to ask. Critical thinking, media literacy, data literacy, numeracy and graphicacy (the ability to interpret graphics, charts and maps) are vital in today's society. Work should always start with the question: Does this information make sense? Is this plausible? These should always be asked alongside the traditional: who, where, what, when, why and how? questions of storytelling.
- The secondary tool to get started in data-driven journalism is an understanding of spreadsheets and basic maths. Other tools are of course available but not necessary.
- There needs to be more extensive structured training on the new tools available to storytellers.
- Newsroom editors would also benefit from training in data-driven journalism so that they can better understand the craft and the value that it can bring to the team and the output.
- National audiovisual archives provide an untapped resource for both data journalists and for debunking mis- and disinformation. The archives can reveal unseen stories and provide context, background and depth to enrich news reports.
- While AI-powered tools are causing disruption in the field of news, they also provide new and exciting opportunities to generate high-value stories.
- There is concern over who controls the foundation models on which many Large Language Models, which are used to train artificial intelligence, are based. Without knowing how the LLM is constructed, there is a risk of introducing bias into tool.
- Media and audiovisual archives can use from Large Language Models to categorise and maximise use of their rich resource but it would be more prudent to build their own.

2. Introduction

Telling stories is universal. It has been part of our DNA since our ancestors were cave dwellers. Prehistoric drawings dating back tens of thousands of years have been discovered in caves and on rock faces in almost every part of the world, from the Chauvet cave (France) to Sulawesi (Indonesia), from Bhimbetka (India) to Serra da Capivara (Brazil), from the Apollo 11 Cave (Namibia) to Blombos in South Africa, to name just a few. They are a testament to our impulse to document the world around us.

> "It's story that makes us human. Recent research suggests language evolved principally to swap 'social information' back when we were living in Stone Age tribes. In other words, we'd gossip. We'd tell tales about the moral rights and wrongs of other people, punish the bad behaviour, reward the good, and thereby keep everyone cooperating and the tribe in check. Stories about people being heroic or villainous, and the emotions of joy and outrage they triggered, were crucial to human survival." (Storr, 2020, p. 2)

Over the millennia, storytelling has evolved to take a myriad of different forms from pictorial to oral, the written word to music, dance to recorded sound and images: poems, myths, fables, novels, plays, photographs or films, the list is almost endless. Some fear however, that today's digital age is impacting the age-old tradition.

> "Now, with 24/7 news cycles, hours spent tweeting or updating Facebook pages with daily minutia, and endless reality television shows, the full power of storytelling - its contextual beauty and majestic ability to move us - is on the wane.
>
> What this means is that today's children may know the facts but not the context in which things happen. As they are no longer being shaped by a storytelling world, they seem to lack the will to dig deeper, preferring to surf in the immediate." (Buster, 2018, pp. 11-12)

While the internet and constant connectivity has the potential to divide attention, the digital age and computing has also ushered in a new form of storytelling. The ability to gather vast amounts of data on any given topic, to organise it into spreadsheets and to use software to analyse the data, draw out trends or contrasts, has enabled storytellers to see previously hidden connections, to tell stories differently and bring depth that goes beyond text, photo or videos in isolation. Combining data analysis with storytelling can bring greater understanding to or shine a new light on familiar topics. Visualisation techniques can present information in a new and engaging way for that story to be told.

For centuries, stories have usually always contained a grain of truth, a lesson from which humans can learn and advance.

> "All we have had from our caveman campfire days until relatively recently (in historical terms) was the oral tradition. The wisdom of the ages has been handed down by the shamans, medicine men, or griots from tribal cultures the world over via folklore, fairytales,

> myths and legends. This is how each generation was psychologically prepared for their future - to be ready, to know that they would not only be able to survive, but that they would thrive amid life's inevitable adversities." (Buster, 2018, p. 11)

However, with an increasing number of people manipulating facts and narratives for their own benefit or for a specific purpose, how can we identify factual stories that are, in reality, false? The rapid spread of misinformation across social media and other outlets has been identified as one of the biggest threats to democracy today.

Could there even be a link between the explosion of open-source databases and some of the misinformation that circulates through social media?

For WP2, both Agence France-Presse (AFP) and Inholland University of Applied Sciences (INH) have taken the lead in researching and compiling the following *State of the Art Update* report in data-driven journalism and storytelling, and digital verification, as well as in conducting a *Needs Analysis (D2.1)* regarding the contents and didactical form of a course to train students in media and creative industries courses in working with data. Additionally, a *State of the Art of Data-driven Work in the Creative and Media Industries* (D2.3) explores how data and AI is impacting music, film and other creative industries. This report will give students an introduction to these vast topics and an understanding of the importance of integrating these skills into their professional and everyday lives. This work also offers insight for educational institutions, media organisations and even the general public as to the role that data analysis and debunking misinformation can play in contributing to a more equitable and balanced society and how we can all be actors in this new form of storytelling.

3. Scope of this Report

This report is divided into five sections. The first section focuses on topics and findings related to data, the second section looks at digital verification, more commonly known as fact-checking, while the third looks at the role of artificial intelligence in data journalism, misinformation and disinformation and more widely in the newsroom, the fourth examines the role of data in national audiovisual archives and the final section provides an overview of EU policy towards disinformation and artificial intelligence.

There is an extensive body of published work going back decades that explains how to work with data: how to collect and process it, how to extract meaning and create visualisations that are informative and impactful. The important role that mis- and disinformation now plays in democracies around the world has also led to a large number of published works and guidelines on the phenomenon so to detail all the means and methods here is beyond the scope of this report. A list of suggested reading, references and resources can be found at the end of this report for those interested in exploring the topics further.

This report will therefore focus on why media students, future and already practising journalists and creatives should learn to work with data, fact-checking and artificial intelligence, how to teach it, hurdles to overcome, some of the recommendations and some of the pitfalls of navigating a world of data processing including spreadsheets, coding, charts, and graphics.

4. Methodology

This chapter offers a brief overview of the research questions and instruments of the field research.

4.1. Main Research Questions

Data science and storytelling with data are both vast fields of work so it was necessary to define exactly where the MediaNumeric consortium had to direct its focus. INH and AFP, therefore, hosted several brainstorming sessions with its partners to discuss, debate and challenge each other on what we felt were the core areas to examine. Together we identified avenues for exploration, people to interview, literature to read and topics to develop.

From these sessions, we formulated the following topical questions:

- What kind of data is available? Where is it? How can you work with data collection holders?
- How to collect, clean, organise and work with data?
- How to develop a critical analysis of data? How to ask the right questions from the data? What are the ethics in using data?
- How to identify a story from both small and large data sets and bring it to life?
- When the story is revealed, how can you use the data to create compelling content? How can you combine other forms of communication (text, photo, video, graphics, animation, etc.) to create impactful work?
- How to harness technology to enrich the storytelling process and experience?
- Are there ethics to storytelling? How do you approach stories with a critical eye? How important is media/story literacy?
- What is disinformation and misinformation? What are its consequences?
- What methods can you use to discover whether a story is true, false or has been manipulated?
- What kind of technology is used to generate misinformation and disinformation to spread throughout social media networks and what technology can be used to debunk those stories?

For the final iteration of this report, AFP held several brainstorming sessions with its partners to discuss areas that merited further study, topics that had emerged over the course of the MediaNumeric programme and recent technological developments which, while not included in the MediaNumeric programme, impact the area of data journalism and fact-checking today and will continue to impact society in the future.

At the core of the project is the following question: **How do you take the age-old tradition of storytelling that is fundamental to human nature and integrate 21st century technology, knowledge and skills?**

4.2. Procedure

As documented in the project proposal, two research methods were chosen for this study — desk research and expert interviews.

4.2.1. Interview Phase

The consortium *(Appendix 1)*, which brings together educational institutions (INA, INH, SWPS), national archives (NISV, INA, FINA - in the initial phase of the project) and industry professionals (AFP, Storytek, CC),[1] together decided on people to be interviewed, including 15 people from the MediaNumeric Stakeholder Board. A total of 56 interviews were carried out over a period of two months. Consortium members were given freedom to choose the most appropriate professionals to question and the interviewees reflect the consortium's various specialities. They come from the world of higher education, journalism and other forms of storytelling, archival institutes, technical and creatives. The audio was transcribed in full using a transcription tool internal to AFP. The interviews were then checked by the interviewers, translated into English if necessary, and collected in a shared document folder. These interviews provided context and content which, along with extensive desk-based research, provided the material to bring together this report.

For the final iteration of this report, further interviews were carried out with lecturers and panellists who participated in the three in-person training sessions in Paris, The Hague and Warsaw, as well as the MediaNumeric final event conference in Warsaw that drew in industry experts. Other insights were drawn from the consortium's own experience and desk research over the course of the three-year project.

4.2.2. Interviews & Selection Procedure

INH and AFP held a series of online brainstorming sessions with all MediaNumeric partners. The project members were asked to complete a form and share their ideas (*Appendix II*). These were reviewed and summarised into one integrated document (*Appendix III*). After further discussion, the designated interviewers settled on a simple and two-fold set of open questions (*Appendix III*). The names of members of the Stakeholder Board were divided among the partners for an interview. The rest of the interviewees were selected by the partner organisations themselves. They approached potential interviewees based on their experience, affiliation, and their relevance for the objectives of MediaNumeric. The basic principle of the selection procedure was to ensure a balance between the experts from the relevant fields — creative industries, mass media, data science and archival work. The list of experts who were interviewed can be found in the list of references at the end of this document.

As well as answering questions, the interviewees provided their own list of tools, literature and other resources which contributed information to the resources section at the bottom of this report.

[1] See Appendix I for full information about MediaNumeric consortium partners.

The interviews were collated together, from which it was possible to compile perspectives and experiences within the fields of media, education, technology and archiving. The results of that compilation follows below.

5. Data

5.1. What is Data?

There is a common misconception among people who do not work closely with data that data revolves around numbers. While numbers do play a dominant role in data-driven journalism, the field stretches as far as any one person's imagination. There are databases that categorise LGBTQ+ art and artefacts[2] or the lines of the hit musical Hamilton,[3] projects that quantify a break-up[4] or analyse similarities between the way that African-Americans play the card game Spades[5] and the way they live their lives. In short, data is simply any kind of information that has been organised, categorised into units and can be quantified.

> "The word "data" means "given" in Latin, in the sense of a "fact". It became the title of a classic work by Euclid, in which he explains geometry from what is known or can be shown to be known. Today data refers to a description of something that allows it to be recorded, analysed, and reorganised. … To datafy a phenomenon is to put it in a quantified format so that it can be tabulated and analysed." (Mayer-Schönberger & Cukier, 2013, p. 78)

5.2. Brief History of Data

The use of data dates back to the moment that man was first able to count. Some of the earliest examples of humans collecting and storing data are notched sticks known as counting or tally sticks. The Ishango Bone, which is exhibited at the Museum of Natural Sciences in Brussels, Belgium, is believed to be one of the oldest mathematical artefacts in existence.[6] The animal bone dating back to between 18,000 BC and 20,000 BC was discovered during archaeological excavations during the 1950s in northeast Democratic Republic of Congo. The handle is carved with 168 parallel marks[7] positioned in groups of parallel lines and split into columns on three sides of the handle. The archaeologist, Belgian Jean de Heinzelin, speculated that the group of lines represented numbers and that the bone was proof of advanced mathematics. Mathematicians have pored over this artefact ever since to decipher its meaning.

---

[2] *Selected Archival Objects & Artworks*. (n.d.). GLBT Historical Society. https://www.glbthistory.org/objects-artwork-selections

[3] Wu, S. (n.d.). *Visualizing Hamilton*. Shirley Wu Studio/. https://shirleywu.studio/hamilton-talk/#/intro

[4] *Quantified Breakup*. (2013, December 31). Quantified Breakup. https://quantifiedbreakup.tumblr.com/

[5] Hickmon, G. (2021, August). *How you play Spades is how you play life*. The Pudding. https://pudding.cool/2021/08/spades/

[6] Swetz, F. (n.d.). *Mathematical Treasure: Ishango Bone | Mathematical Association of America*. Mathematical Association of America. https://www.maa.org/press/periodicals/convergence/mathematical-treasure-ishango-bone

[7] *Have you heard of Ishango?* (n.d.). Royal Belgian Institute of Natural Sciences. https://www.naturalsciences.be/sites/default/files/Discover%20Ishango.pdf

Figure 1. The four sides of the Ishango Bone with the quartz mounted to one end of the bone. [Screenshot]. Royal Belgian Institute of Natural Sciences.[8]

As well as bones, our ancestors also used small stones to measure, log and record. The word calculate comes from the Latin word calculus meaning a small pebble. Humans took the concept of pebbles and put them in counting boards and counting frames, using beads to construct abaci.

With the advent of writing, data could take on a new form and recorded information started to fill libraries around the world. Surveyors were dispatched to make maps of cities and countries far and wide. One of the most famous early exercises in data gathering was the Domesday Book,[9] a detailed survey and valuation of landed property in England at the end of the 11th century. Carried out in 1086 on the orders of William the Conqueror, the Domesday painstakingly recorded details of 13,418 places including the name of each manor, who owned it in the time of King Edward in 1066, who owned it at the time of the survey, the size of the land, how the land had changed, how many people worked on the land, the number of livestock, and much, much more. It was noted by an observer of the survey that "there was no single hide nor a yard of land, nor indeed one ox nor one cow nor one pig which was left out" (Johnson, n.d.).

At the end of the first millennium, the Hindu-Arabic numeral system composed of the ten symbols 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9 was introduced to Europe and over the following centuries, they gradually replaced the Roman numeral system that had originated in ancient Rome. The Hindu-Arabic numeral system enabled much more complex maths to be undertaken.

[8] Figure 1: *Have you heard of Ishango?* (n.d.). Royal Belgian Institute of Natural Sciences. https://www.naturalsciences.be/sites/default/files/Discover%20Ishango.pdf
[9] The National Archives. (n.d.). *Domesday Book*. https://www.nationalarchives.gov.uk/help-with-your-research/research-guides/domesday-book/

One of the first recorded exercises in statistical analysis was published in 1662 by John Graunt, a haberdasher in London by trade but who is regarded by many historians as the founder of the science of demography. Graunt studied 50 years of records held by London parishes and classified the deaths according to cause of death and gender. In his book *Natural and Political Observations Made Upon the Bills of Mortality* (Graunt, 1660-1674), Graunt reached rudimentary conclusions about the mortality and morbidity of certain diseases. He was highly sceptical of the number of deaths recorded as due to the plague and he speculated about the reasons for these misclassifications.

Towards the end of the 19th century, the US Census Bureau estimated that it would take approximately eight years to process the data collected during the 1880 census and a whole decade to process the data generated by the 1890 census. The data would be outdated before it could even be released. In order to process the information in a timely manner, a young engineer named Herman Hollerith devised punch cards to record information. His machine, the Hollerith Tabulating Machine, reduced 10 years of work to just three months. The company he founded went on later to become IBM.

At about the same time, the English nurse, social reformer and statistician Florence Nightingale was caring for soldiers wounded in the conflict between the Ottoman and Russian empires, a conflict that would later become known as the Crimean War. She kept thorough records of all patients and events and on her return to England in 1886, she analysed the data in collaboration with statisticians, organising it in new ways to present striking visualisations. Her most famous chart, known as a rose chart or polar-area diagram, shows the cause of deaths among the ranks of the British army in military hospitals and camps during the war in Crimea. To her horror, Nightingale discovered that more soldiers died from disease than died from their battle wounds and that the diseases had been able to spread due to poor sanitation and poor ventilation in the hospitals (Cairo, 2019, Don't Lie to Yourself (or to Others) with Charts). Her charts and diagrams were among the first that were specifically designed to persuade people of the need for change.

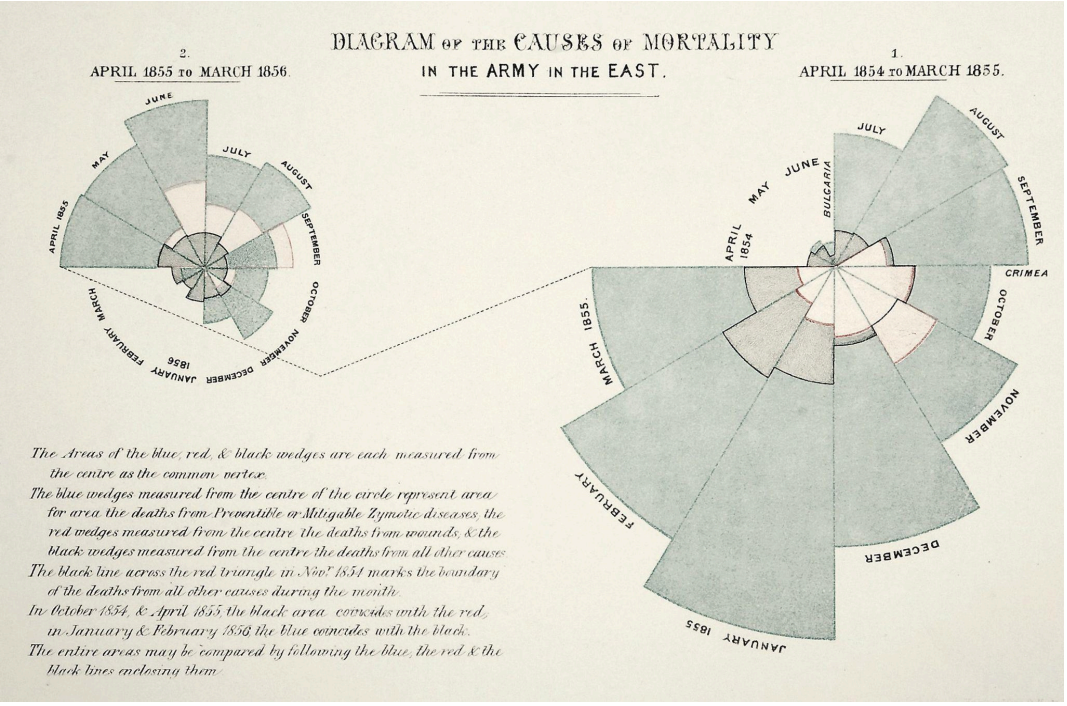


Figure 2. Diagram of the causes of mortality in the Army in the East. (n.d.). [Photograph]. National Library of Medicine.[10]

In the first half of the 20th century, new discoveries enabled data to be stored to magnetic tape and counting operations were able to analyse large volumes of data. In 1989-1990, computer scientist Tim Berners-Lee created hyperlinks, hypertext and enabled data sharing worldwide. His work resulted in the World Wide Web and the basis of the internet as it is known today. The arrival of Google and Google search functions in 1998, cloud storage and artificial intelligence systems has accelerated the compilation, processing and analysis of data at an increasingly rapid pace, leading to an era of what is known as ”big data”.

> “Half a century after computers entered mainstream society, the data has begun to accumulate to the point where something new and special is taking place. Not only is the world awash with more information than ever before, but that information is growing faster. The change of scale has led to a change of state. The quantitative change has led to a qualitative one,” (Mayer-Schönberger & Cukier, 2013, p. 6)

Mayer-Schönberger and Cukier specify that there is no rigorous definition of big data but over the years the concept has become known by a series of V's. Big data is:[11] [12]

[10] Figure 2: *Diagram of the causes of mortality in the Army in the East - Digital Collections - National Library of Medicine*. (n.d.). National Library of Medicine. https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101598842-img
[11] Arockia, P., Varnekha, S., & Veneshia, K. (2017). The 17 V’s Of Big Data. *International Research Journal of Engineering and Technology (IRJET)*, *04*(09), 330–331. https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf
[12] Serokell, S. (2021, January 8). *What Is Big Data?* Medium. https://ai.plainenglish.io/what-is-big-data-646d5e5bedc3

The "V"	Definition
Volume	Size of data collected: massive amount of data collected and stored, calculated in terabytes, petabytes, exabytes and zettabytes
Velocity	Speed of data: how quickly data is generated, travels and is processed, data moves as a constant stream of information
Value	Importance of data: business value to be gained from the data
Variety	Type of data: different types of data sources and file types and whether the data is structured (already organised, formatted and convenient to work with), unstructured (unorganised, unformatted, or file types that cannot be organised into rows and columns, such as photos or videos) or semi-structured (does not conform to a formal structure of data)
Veracity	Quality of data: captured data should be accurate and trustworthy.
Variability	Changeability of data: data whose meaning is constantly changing.

Table 1. Description of the six V's of big data.

There is some debate about how many V's should be used to define big data. Some stop at 5, other sources list the six above, others suggest 15 different characteristics starting with V adding validity, volatility, viability, visualisation, virality, viscosity, vocabulary, vagueness and venue while others suggest 17 or more.

The relentless stream of data has led to a whole new area of work known as data science. This encompasses specialist fields such as data mining and statistical analysis, data engineering and data warehousing, database management and data analytics. The list is extensive.

> "The era of big data challenges the way we live and interact with the world. Most strikingly, society will need to shed some of its obsession for causality in exchange for the simple correlations: not knowing *why* but only *what*. This overturns centuries of established practices and challenges our most basic understanding of how to make decisions and comprehend reality." (Mayer-Schönberger, 2013, pp. 6-7)

Big data is usually processed by internet companies, social media platforms, banking and finance, and any other entities that have massive computing power at their disposal. This report will focus on data that can be contained within spreadsheets and processed by easily accessible software. And while journalists start with the "*what*", they are still equally interested in the "*why*".

Journalists have always used some form of data collection and analysis in their reporting but it was in the 2000s that access to data online began to play a role in large-scale investigations, including the UK parliamentary expenses scandal (2009), the WikiLeaks diplomatic cables (2010), the

offshore leaks (2012-2013), Luxembourg Leaks (2014), Swiss Leaks (2015), Panama Papers (2016), Paradise Papers (2017), FinCEN Files (2020), OpenLux (2021) and the Pandora Papers (2021). Further information about these investigations can be found in parts 5.4.2 and 5.5.

5.3. Fields of Data

There are many fields of data science. The **three fields most involved in data-driven storytelling are data collection, data exploration and visualisation**. Each field requires a different level of knowledge of maths, coding, computer sciences or visualisation techniques. An individual's knowledge and skill base in each field dictates how independent they may be when carrying out certain tasks. However, all experts stress that it is in no way necessary to be proficient in all fields. An understanding of basic maths plus familiarity with spreadsheets, is enough to get started.

5.4. What is Data-Driven Storytelling & Why Is It Important?

Data-driven journalism and storytelling is essentially any story that rests in some part on the analysis of a data set (Cairo & Rogers, 2021, April 27). Many stories contain visual elements, such as charts, maps, or graphics to illustrate the data. However, not all data-driven journalism includes data visualisation. Articles published by *Le Monde* newspaper, for example, about the 2021 OpenLux investigation into corporation tax practices in Luxembourg contained few visualisations even though the entire investigation was based on data (Kelly, 2021, No. 38).

> *"For me, data journalism is essentially anything that has to do with journalism and data, that's essentially it. So if you use data to do investigative reporting, you are doing data journalism. If you are visualizing data to present it to the public, you're doing data journalism. If you are doing algorithmic accountability on ethics in the context of a journalistic endeavour, you're doing data journalism. So anything that has to do with computational methods and quantitative data in relation to journalism, I think that we could label it as data journalism."*
>
> - *Alberto Cairo, Knight Chair in Visual Journalism at the School of Communication of the University of Miami (Cairo & Rogers, 2021, April 27).*

The ability to analyse data sets enables journalists to verify for themselves information that is provided to them by a source, an organisation or a government. These organisations may be promoting a specific agenda and have selected a certain piece of the data that supports their goal. Instead of reporting the supplied figures and assuming them to be correct, journalists can dive into the data, test it, come to their own conclusions, and even find different or additional stories in the data. It provides more depth and can illuminate hidden aspects of a story.

5.4.1. Engagement of Audiences and Monetisation

With news and information so widespread on the internet and advertising revenue spread across so many different online and offline spaces, media companies are all struggling with revenue and distribution. It can be difficult to monetise general news reports which can be easily copied across competing news organisations. Data-driven stories are infinitely more complex and detailed and as such, although competing news organisations are able to report on the findings, they cannot replicate at speed the intricacies of the investigation. As a result, the stories are much more copy-proof. They are popular with readers and provide an opportunity to draw in subscribers or increase advertising revenue. The most popular content on *The Washington Post* is usually data visualisations (Cairo, 2021, MediaNumeric interview). The newspaper "disclosed last year [eds: 2020] that out of the seven most popular stories ever published by *The Washington Post* online, six of them were stories that had been produced by the graphics department," said Cairo (Cairo, 2021, MediaNumeric interview). For years, the most popular story published by the New York Times was a data visualisation while one quarter to one third of traffic at ProPublica happens through interactive databases or data visualisations (Cairo, 2021, MediaNumeric interview).

For some media organisations in the world, this trend is contributing to the creation of a new business model as Aron Pilhofer, the James B. Steele Chair in Journalism Innovation at Temple University explains:

> "I can tell you definitively that as we are shifting away from scale and reach and more toward reader revenue, membership, that people pay for exclusive things. People pay for content that you can't get anywhere else. And in a world now where exclusives are exclusive for about a nanosecond before someone aggregates them and they're immediately all around the web, the thing that can remain exclusive, the thing that's impossible to aggregate and rewrite quickly is data journalism, a great visualisation, great content**.** This is why the New York Times is doubling, tripling, quadrupling down. This is a business decision. This isn't because they love making pretty pictures on the Internet. They're making a business decision that is key to their future." (Cairo & Rogers, 2021, April 27)

This is not the case everywhere though. In France, for example, data journalism has still not become as mainstream as in many of the major Anglophone markets, according to Maxime Vaudano, who works at Les Décodeurs at *Le Monde* newspaper.

"There is little recognition and little desire on the part of the editorial offices, which are globally, aside from data, in a very conservative movement and retreating from innovation and from economic models with a renewed focus on paying models.

Due to economic pressure and the concentration on an exclusive subscription model, newsrooms are giving journalists much less 'slack' with which to experiment on new things. Whether it be right or wrong to do so, newsrooms believe that subscribing internet users are not necessarily the same people who consume data journalism productions." (Vaudano, 2021, MediaNumeric interview).

5.4.2. Impact

Data-driven stories have the ability to hold institutions, companies, governments, and powerful people to account, as demonstrated by the list of investigations that follows.

> *"You want to be able to hold powerful people to account for the decisions they make that affect everybody in society and believe me the power of data is going to be what the powerful people…the people that control the data are going to be in power. And so your ability to understand that and to interrogate that is absolutely fundamental to your purpose as a journalist."*
>
> - *Peter Rippon, editor of BBC Online Archive, UK (MediaNumeric interview)*

In 2002, the *Boston Globe*'s Spotlight investigative team published a series of explosive reports[13] into the widespread abuse of children by Catholic priests and the local archdiocese's systemic efforts to cover up the crimes over decades. The scandal rocked the Roman Catholic Church to its

[13] Boston Globe. (2002, January 6). *Church allowed abuse by priest for years*. BostonGlobe.Com. https://www.bostonglobe.com/news/special-reports/2002/01/06/church-allowed-abuse-priest-for-years/cSHfGkTIrAT25qKGvBuDNM/story.html ; Boston Globe. (2002b, January 7). *Geoghan preferred preying on poorer children*. BostonGlobe.Com.https://www3.bostonglobe.com/news/special-reports/2002/01/07/geoghan-preferred-preying-poorer-children/69DE1kOuETjphwmIBcgzCM/story.html?arc404=true ; Boston Globe. (2002c, January 31). *Scores of priests involved in sex abuse cases*. BostonGlobe.Com. https://www.bostonglobe.com/news/special-reports/2002/01/31/scores-priests-involved-sex-abuse-cases/kmRm7JtqBdEZ8UF0ucR16L/story.html
Boston Globe. (2002d, December 24). *Church cloaked in culture of silence*. BostonGlobe.Com. https://www.bostonglobe.com/news/special-reports/2002/02/24/church-cloaked-culture-silence/88cLKuodvSiHjvg0dfz24L/story.html

core and spawned similar investigations around the world. The ripple effect of the investigation continues to this day.

"The Spotlight project opened the floodgates on clergy sexual abuse, locally, regionally, nationally, and internationally," Boston lawyer Mitchell Garabedian told the *Boston Globe* in a follow-up report[14] 20 years after the story broke. (Kahn & Damiano, 2021).

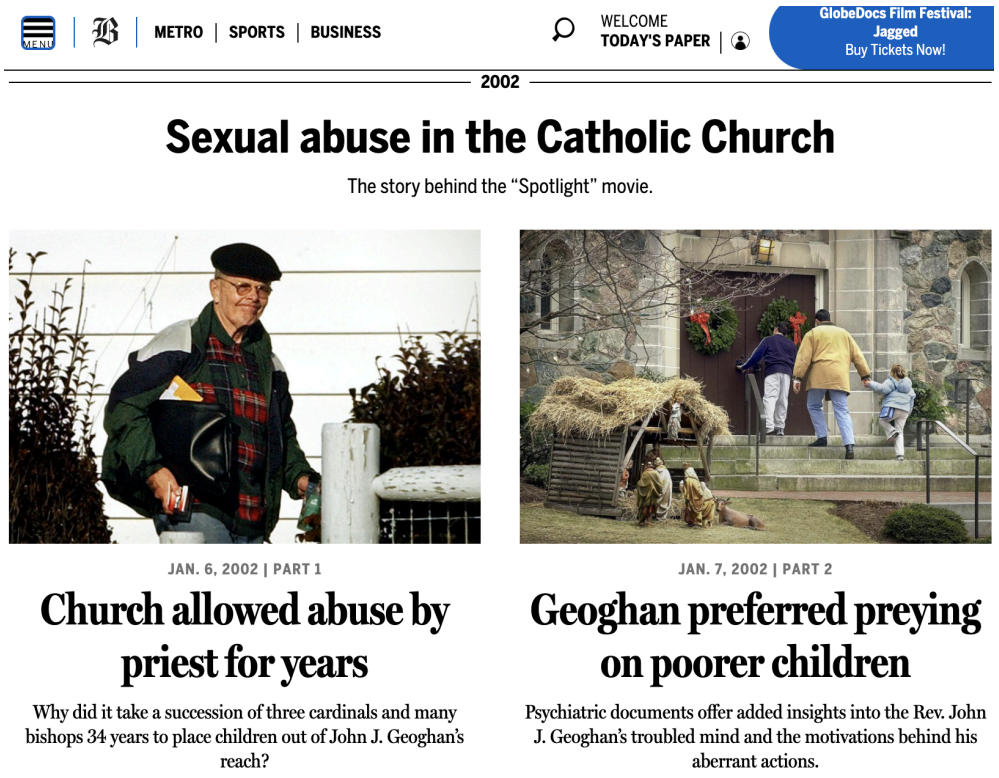


Figure 3. Screenshot of front pages of Boston Globe's coverage of church child abuse scandal.[15]

While the sexual abuse by priests was already known, it was only when the team manually combed the church's printed directories and created a spreadsheet to track how and when priests were reassigned to different roles that the journalists realised and, more importantly, were able to prove that church authorities knew about the abuse and actively covered it up.

"Spotlight's impact was so far-reaching, it's hard to measure in concrete terms," [clergy sexual abuse survivor Phil] Saviano said. In his view, it also "led to a reckoning with the Boy Scouts, at Penn State, even the #MeToo movement as it gathered steam." Without Spotlight's reporting, he believes, Pope Francis would never have convened the 2019 Vatican summit on clergy abuse, during which the pontiff urged bishops to "listen to the cry of the children who ask for justice." (Kahn & Damiano, 2021).

[14] Kahn, J. P., & Damiano, M. (2021, September 22). *'They knew and they let it happen': Uncovering child abuse in the Catholic Church*. Boston Globe. https://www.bostonglobe.com/2021/09/22/magazine/they-knew-they-let-it-happen-uncovering-child-abuse-catholic-church/

[15] Figure 3: Boston Globe. (2002, January 6). *Sexual abuse in the Catholic Church* [Screenshot]. Boston Globe. https://www.bostonglobe.com/metro/investigations/spotlight/?p1=Article_Inline_Text_Link

The investigation focussed on a very specific piece of data in the Boston area of Massachusetts, United States, notably priests changing parish, and yet the ramifications of the investigation were global and have continued for two decades. The revelations led to a closer examination of sexual abuse in the Roman Catholic Church in countries around the world, specifically how the church hierarchy actively protected alleged abusers. A hitherto taboo subject exploded into the open and survivors of abuse found the courage to speak out. It led to convictions of high-ranking clergy leaders and even former pope Benedict XVI was accused of failing to act.[16]

More recent investigations have had equally far-reaching consequences, such as the reports detailed below conducted by the International Consortium of Investigative Journalists (ICIJ) and their partner media organisations.

Figure 4. Screenshot of front page of ICIJ Panama Papers investigation.[17]

Thanks to a leak of more than 11.5 million financial and legal records, journalists reporting on The Panama Papers[18] revealed in 2016 how offshore tax havens created a system to enable tax evasion, corruption and fraud. Tax lawyers were jailed, Iceland's prime minister and Spain's industry minister were forced to step down and countries around the world worked to tighten money

[16] BBC News. (2021, December 8). *Ex-Pope admits errors in handling of abuse cases*. https://www.bbc.com/news/world-europe-60305844

[17] Figure 4: ICIJ. (2017, January 31). *The Panama Papers: Exposing the Rogue Offshore Finance Industry* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/panama-papers/

[18] ICIJ (2021, April 3). *The Panama Papers: Exposing the Rogue Offshore Finance Industry*. ICIJ. https://www.ICIJ.org/investigations/panama-papers/

laundering laws. Former South African President Thabo Mbeki called the revelations "a massive blow to financial secrecy."[19] (Fitzgibbon & Hudson, 2021).

Another leak of 13.4 million confidential files dubbed the The Paradise Papers[20] shone a light on one of the world's most prestigious offshore law firms, including the interests and activities of more than 120 politicians and world leaders, including Queen Elizabeth II, and 13 advisers, major donors and members of then US President Donald J. Trump's administration. The revelations pushed more than 130 governments around the world to sign a global agreement on minimum corporate tax rate.[21]

Figure 5. Screenshot of front page of ICIJ Paradise Papers investigation.[22]

The first global investigation into the medical device industry, known as the Implant Files,[23] tracked the harm caused by medical devices, such as breast, buttock and pelvic mesh implants, that had been tested inadequately or not at all. Following the findings, governments around the world updated safety reviews, cut down on the use of some implants and took steps to reform and regulate the industry.

Figure 6. Screenshot of front page of ICIJ Implant Files investigation.[24]

[19] Fitzgibbon, W. (2021, April 8). *Five years later, Panama Papers still having a big impact*. ICIJ. https://www.ICIJ.org/investigations/panama-papers/five-years-later-panama-papers-still-having-a-big-impact/
[20] ICIJ, (2020, July 15). *Paradise Papers: Secrets of the Global Elite*. ICIJ. https://www.ICIJ.org/investigations/paradise-papers/
[21] ICIJ, (2021, October 11). *136 countries agree to global minimum tax for corporations in 'historic' OECD deal*. ICIJ. https://www.ICIJ.org/investigations/paradise-papers/136-countries-agree-to-global-minimum-tax-for-corporations-in-historic-oecd-deal/
[22] Figure 5: ICIJ. (2017b, November 5). *Paradise Papers: Secrets of the Global Elite* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/paradise-papers/
[23] ICIJ. (2020, September 4). *Implant Files*. https://www.ICIJ.org/investigations/implant-files/
[24] Figure 6: ICIJ. (2018, November 25). *Implant Files* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/implant-files/

The Pandora Papers[25] is the ICIJ biggest investigation to date. Published in 2021, the data leak revealed how 14 offshore service providers — law firms, wealth management advisors and corporate formation agencies — helped the world's wealthy set up companies in low- or no-tax jurisdictions in order to avoid paying tax.

Figure 7. Screenshot of front page of ICIJ Pandora Papers investigation.[26]

In the investigation by the *Boston Globe*, journalists had to scour the Church's paper records to gather the information necessary to their investigation. The collection of data was manual and required the journalists to enter the information into spreadsheets themselves. The scope of their investigation was limited to the city of Boston so they were able to keep the investigation within the Spotlight team only.

In the ICIJ investigations, the content arrived as digital files. The dumps were huge with millions of documents to examine. The data was collated, organised and cleaned by the ICIJ technical team but as the information impacted almost every corner of the globe, the ICIJ invited partner media organisations to examine the files and use their local knowledge to find the stories hidden within it. The data trail was just the start of the investigation. The data needed to be tested and verified with the basic principles in information gathering: Who? What? When? Where? Why? and How? By collaborating with a wider network of journalists, the data dump was able to yield a much broader picture of wrongdoing that was replicated the world over.

The ICIJ provides extensive coverage of their special reports on their website www.ICIJ.org/investigations and often provides information about how the investigation was led. In a *Conversations With Data* podcast with host Tara Kelly, the ICIJ's Chief Technical Officer Pierre Romera goes behind the scenes of the Pandora Papers and explains how they worked with hundreds of journalists around the world on the wide-ranging and complex investigation (Kelly, 2021, Episode 38).

The examples above from the ICIJ are wide-ranging global investigations that involved many partners across the world that led to real-world action and legislation. There are countless examples of smaller national, regional or local data-driven investigations that equally led to concrete and beneficial action.

[25] ICIJ. (2021, December 20). *Pandora Papers*. https://www.ICIJ.org/investigations/pandora-papers/
[26] Figure 7: ICIJ. (2021, October 3). *Pandora Papers* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/pandora-papers/

These types of stories also inspire staff loyalty, engagement and a sense of pride among the creators. All of the journalists working in the field say they believe they are fulfilling their public service mandate to investigate and to inform when they work on these types of projects, as Scott Klein who leads the teams at the intersection of journalism and technology at ProPublica explains:

> "We're a non-profit organisation. We're not just a newsroom that just seeks to publish stories for our readers, we want to have impact in the real world. It is very much our mission to make the world a better place through our journalism. So on some level it's straightforward for us to look at a story pitch and to understand whether it makes sense for ProPublica because we're about impact and we see in stories chances for impact." (Cairo & Rogers, 2020, Dec. 17)

Mike Rezendes, one of the *Boston Globe* reporters who investigated the Catholic Church paedophilia scandal concurs.

> "Well, it may sound a little trite, but in fact, it's not trite at all. It's very serious and it's very true. This is how I make my contribution to society. This is how I make sure I leave the world a little better off than it was when I arrived. I am committed to the work and there really is nothing else in journalism I would rather do." (Rothman, & Miller, 2016)

5.6. Data Transparency

Part of the power and impact of data-driven storytelling comes from the openness of many of the data sets and the openness of the reporting.

In an interview with ABC News at the time of the release of the Hollywood film *Spotlight* which recounts the *Boston Globe* investigation into the Roman Catholic child sex abuse scandal, Rezendes said:

> "We published our stories right at the dawn of the internet era. When we published day 1 and day 2, they were based on the church's own internal documents and we actually put those documents up on line. Today that is standard procedure but back then, it seemed pretty novel.
>
> It showed people that our story was irrefutable, it showed people that everything in the story was nailed to the ground, everything in the story was based on the church's own internal documents." (Rothman, & Miller, 2016)

These two aspects – that the story was based on the church documents and that the *Boston Globe* published those documents and their process to analyse them – gave the story a level of impact that could in no way have been obtained if it had been solely based on survivor testimony or

whistleblowers. The church could not refute the facts detailed in their own documents whereas they may have attempted to contest a personal account.

In February 2021, a consortium led by *Le Monde* newspaper published a series of articles[27] into the tax advantages offered by the financial centre in Luxembourg.[28] Using open source data, the 11-strong team, which included *Suddeutsche Zeitung* in Germany, *Le Soir* in Belgium, McClatchy in the United States, *Woxx* in Luxembourg, IrpiMedia in Italy, and the OCCRP Consortium of investigative journalists, compiled a huge database of the beneficiaries of 124,000 commercial companies registered in Luxembourg. Information included the companies' true owners and 3.3 million administrative acts and financial reports.

From these records, reporters were able to establish that Luxembourg effectively acted as an offshore tax haven that facilitated tax avoidance but one that was located in the heart of the European Union. According to reporters who worked on the project, *Le Monde* journalist Maxime Vaudano and OCCRP editor Antonio Baquero, the Luxembourg authorities contested the journalists' line of questioning at the start of the investigation but by the end, they had to accept the evidence that the reporters had uncovered.

> "Beyond Luxembourg itself, the project itself was a good proof of concept of transparency because all the information that we started with was open source information that was publicly available but poorly accessible. What we did as journalists was to access this, to filter it for the public because it's impossible for the public to go through hundreds of thousands of companies. But in the end, it was just proof that transparency works and all that we have been advocating for in the past few years about corporate transparency, about the fact that we have to know who is behind the companies, what the companies are used for," Vaudano said. (Kelly, 2021, No. 35)

The operation to "scrape" the website (collect data from a digital source) to obtain the necessary data, to filter it and to bring understanding to the data, had been very complex. It took a journalism investigation to bring the information together in that way. Vaudano said that it was unlikely that the Luxembourg authorities had realised the extent of the tax fraud being perpetrated by some of the people setting up companies there.

Baquero said that the Luxembourg authorities had grossly inflated the percentage of companies that had declared the ultimate beneficial owner (a measure to fight money laundering) and that

---

[27] Baruch, J., Ferrer, M., Vaudano, M., & Michel, A. (2021, December 9). *OpenLux : the secrets of Luxembourg, a tax haven at the heart of Europe*. *Le Monde*. https://www.lemonde.fr/les-decodeurs/article/2021/02/08/openlux-the-secrets-of-luxembourg-a-tax-haven-at-the-heart-of-europe_6069140_4355770.html

[28] Damgé, M., Michel, A., Vaudano, M., Baruch, J., & Ferrer, M. (2021, March 1). *OpenLux : enquête sur le Luxembourg, coffre-fort de l'Europe*. *Le Monde*. https://www.lemonde.fr/les-decodeurs/visuel/2021/02/08/openlux-enquete-sur-le-luxembourg-coffre-fort-de-l-europe_6069132_4355770.html

this investigation was yet further proof of the necessity for journalists to check the data rather than accept the official figures (Kelly, 2021, Episode 35). This is a practice that must change, said Peter Burger, assistant professor at Leiden University: "Quite a few of them tend to accept, for instance, press releases at face value; will just use one source and not dig into the original publication; lack the skills to judge statistics." (Burger, 2021, MediaNumeric interview)

Working with open source data can even provide a form of protection, particularly to journalists working in areas of the world with poor records of press freedom. Dozens of journalists are killed every year and even more are threatened, injured or forced into exile. Working on issues such as corruption or mismanagement can be dangerous. Using open source data can mitigate the risk.

> "Corruption-related data stories tend to be safer because of the data … often because they are not using leaked data or anonymous sources. They are using publicly available, published data which makes it a little bit harder for the government to come back and criticise them," Eva Constantaras said (Cairo & Rogers, 2021, Sept 28).

Eva Constantaras, a Google Data Journalism Scholar and Fulbright Fellow is specialised in building data journalism teams in countries known as the Global South (a term used to identify lower-income countries). She cited examples of journalists in Kenya and Pakistan whom she trained who both published reports that were critical of government spending. Both reports received a lot of push-back from the authorities but as the reports were based on the government's own figures, the officials backed down fairly quickly.

> "He felt comfortable publishing that story because he was using the Ministry of Health and Ministry of Education, Infrastructure, data and budgets, and reported results. He was able to actually produce a fairly critical story without very much backlash that he would have faced by doing this through anonymous sources or any other kind of more controversial source material." Constantaras said (Cairo & Rogers, 2021, Sept 28).

As well as providing a degree of security for reporters, publicly available data enables journalists to test systems and push for greater transparency in established democracies. It is arguable that the 2021 agreement on a global minimum corporate tax rate would not have come about without the revelations of offshore tax practices revealed in the Panama and Paradise Papers. "I think that having better quality data is useful for anyone, not just for the professionals." (Rendina, 2021, MediaNumeric interview).

5.7. Data Literacy & Graphicacy

Along with the ability to read words and text (literacy) and numbers (numeracy), it is becoming increasingly important in today's society to be able to read, understand and communicate data as information (data literacy) and also to read and interpret infographics (graphicacy).

5.7.1. Data Literacy

There is a common misconception that data is objective. People are more likely to believe you when using a number in your story, says Lindsey Cook, a senior editor, digital storyteller and trainer at the *New York Times*. "That is an enormous power and that is a power that we must make sure that we don't misuse." (Cook, 2021, MediaNumeric interview)

Neuroscientist Daniel Levitin concurs.

> "Statistics, because they are numbers, appear to us to be cold, hard facts. It seems that they represent facts given to us by nature and it's just a matter of finding them. But it's important to remember that *people* gather statistics. People who choose what to count, how to go about counting, which of the resulting numbers they will use to describe and interpret those numbers. Statistics are not facts. They are interpretations." (Levitin, 2017, p. 3)

Many people do not feel confident enough in their own maths abilities to really examine numbers, to question and to test them. However, we should be applying the same critical thinking to numbers and data sets that we do to words and text. When presented with a data set, we should ask: Does this make sense? Is this logical? Data literacy and thinking critically about numbers must be encouraged across all sectors of society, not just journalism.

> "It is easy to lie with statistics and graphs because few people take the time to look under the hood and see how they work. … We – each of us – need to think critically and carefully about the numbers and words we encounter if we want to be successful at work, at play, and in making the most of our lives. This means checking the numbers, the reasoning, and the sources for plausibility and rigor." (Levitin, 2017, pp.ix-x)

Alexandre Léchenet, a data journalist who has worked for many publications in France including *Le Monde* and *Libération*, says that everyone should have basic numeracy skills, an understanding of numbers and data. But they should also ask how the data is constructed, what do the numbers tell us? And just as importantly**, what do the numbers not tell us?** "It is important that students learn to take the data as a source and know how to examine it properly, which does not involve special computer or graphic skills." (Léchenet, 2021, MediaNumeric interview)

Sophie Jehel, lecturer in information and communication sciences, public law, economic and social sciences at the University of Paris 8 Saint Denis, puts it in starker terms.

> "We have been addicted, since the beginning of the Covid crisis, to statistics, but at the same time we have no idea what the value of these figures is or what their margin of error is. This poses ethical problems in the sense that we should, at the same time as giving information, give the margins of error, but this isn't done - all the more so as the data in question guides governments' policy choices and impacts people." (Jehel, 2021, MediaNumeric interview)

5.7.2. Graphicacy

As well as developing skills in data literacy, it is also important to develop the ability to interpret graphics, charts and other visualisations. Humans are first and foremost visual beings. A team of neuroscientists from the Massachusetts Institute of Technology found that the human brain can process entire images that the eye sees for as little as 13 milliseconds (Trafton, 2014). We were able to see and comprehend long before we invented writing and reading. This means that articles with visuals or stories that are made entirely out of visuals are especially compelling. There can be a tendency to look quickly at a data visualisation and think that we have understood it. However this can lead to errors.

> "We see because we have a previous understanding of certain things. Seeing precedes understanding, and this understanding precedes a better, deeper seeing down the road." (Cairo, 2012, Introduction)

Data visualisations are usually more complex than at first glance. They require study to ensure that the scale, the axis and the structure of the visualisation is interpreted correctly. Charts can be manipulated to present information in a way that may be misleading, such as not starting the bar chart scale at zero or creating a break in a bar so that the two bars cannot be easily compared.

> "The persuasiveness of charts has consequences. Very often, charts lie to us because we are prone to lying to ourselves. We humans employ numbers and charts to reinforce our opinions and prejudices, a psychological propensity called the confirmation bias." Cairo says. (Cairo, 2019, Introduction)

Cairo calls the ability to read visualisations "graphicacy," a term coined by geographer William G.V. Balchin in the 1950s. He says that people's ability to interpret graphics, charts, maps and other visualisations has increased considerably over the past few decades but says that he was prompted to write his book *How Charts Lie: Getting Smarter About Visual Information* (2019) because he still sees so many errors.

> "I've seen people misinterpreting graphs systematically. I have seen myself misinterpreting graphics when not applying some of the recommendations that I give in the book, for example, paying attention.
>
> First above all, always internalise the idea that a visualisation is not an illustration. It will never be something that you can interpret in the blink of an eye. A visualisation needs to be approached as if it were an argument or a written text. You need to stop, look at it and then read it. And read it attentively, because if you don't, you will reach the wrong conclusion. Some people find that visualisation should be intuitive very, very quickly. That's wrong. Visualisations can be intuitive in some cases and can be read but in many other cases, they require our attention, they require care." (Cairo, 2021, MediaNumeric interview)

Examples of visualisations that can be misleading at first glance or visualisations that have been constructed in such a way to be misleading can be found in section 5.13.

5.8. Collaboration

Collaboration is vital in the world of data-driven storytelling. It is rare for journalists to be an expert in all fields: data collection, data cleaning, statistical analysis, data visualisation, etc. Each area needs the expertise of the other. For example, a data analyst working on a climate story will need the expertise of a specialist reporter to identify whether the data is "clean" (accurate or plausible) or whether there are figures that need to be investigated and potentially corrected. Reporters need the help of data visualisation specialists in order to transform the information into striking visualisations to draw the reader in and give meaning to the data. "I think it's very important that students have a very broad knowledge of what is possible, kind of a mindset. And then in-depth knowledge of the different specialisations because nobody can do everything as we know." Polish journalist Daniela Kraus said (Kraus, 2021, MediaNumeric interview).

5.8.1. Broader Range of Story Angles

When working on large data sets such as the Panama, Paradise or Pandora Papers, the volume of data is so large that it requires the expertise of many different fields and organisations. The Pandora Papers contained 11.9 million documents with 2.9 terabytes involving companies and people around the world. For one organisation to analyse it all would have taken an impossibly long time and not every story is of interest to every media organisation but every media organisation would certainly find a story that was relevant to their readers and local audience.

In the end, more than 600 journalists from 150 media outlets reaching 117 countries worked on the Pandora Papers.[29] It was the ICIJ's biggest collaboration to date. It allowed the journalists to cover a much more extensive number of stories and different angles. They also had expertise of journalists working in other countries or specific fields who could provide context and background to the data. Everyone was able to contribute to release a much more impactful body of work.

The ICIJ's Chief Technical Officer Pierre Romera said all partners were encouraged to sign up to their philosophy of "radical sharing" (Kelly, 2021, No. 38). Despite news editors' reservations, partners were asked to publish all their findings, leads, testimony, videos and any other information they found in the data on the consortium's secure Global iHub. This allowed partners to collaborate and coordinate extensively but it also enabled the consortium members to publish stories about events in countries where political press or censorship prevented the local media from printing those stories.

[29] ICIJ. (2021a, October 4). *Pandora Papers journalists and media partners*. https://www.ICIJ.org/investigations/pandora-papers/pandora-papers-journalists-and-media-partners/

This type of collaboration can be found across the sector, not just on large-scale projects such as those led by the ICIJ. Alexandre Léchenet participates in the data + local[30] collective, a network of journalists working for regional dailies.

> "The idea is to work with several people on the same database, without being competitors on a local level, to have a multiplicity of processing angles on the same subject and to consequently achieve economies of scale in terms of working time. This network is also characterised by the diversity of journalists' profiles, with people who are self-taught." (Léchenet, 2021, MediaNumeric interview)

He cites an example of journalists from different newspapers investigating together a story about pharmaceutical companies paying doctors. Each journalist analysed the data at their own local level.

> "The end result is interesting, because this collaborative work around data has allowed a rather complex subject to emerge that we likely would not have seen otherwise. The collaboration of journalists from different newsrooms around the same investigation, through data, also brought out things that a single person or newsroom could not have done alone."

5.8.2. Different Specialities

This diversity is cited by specialists time and again. "People have different specialisms, different geographic areas and those can help you produce something together that is much greater than you would be able to do individually," said Simon Rogers (Kelly, 2020, Episode 3).

The Guardian newspaper brought in visualisation experts Nadieh Bremer and Shirley Wu to work on their award-winning story on homeless people in the United States who are given a one-way ticket to states elsewhere in the country.[31]

[30] *data+local - Collectif de datajournalistes locaux*. (n.d.). Data+local. https://collectif-datalocal.github.io/

[31] Outside in America team. (2017, December 20). *Bussed out: how America moves thousands of homeless people around the country*. The Guardian. https://www.theguardian.com/us-news/ng-interactive/2017/dec/20/bussed-out-america-moves-homeless-people-country-study

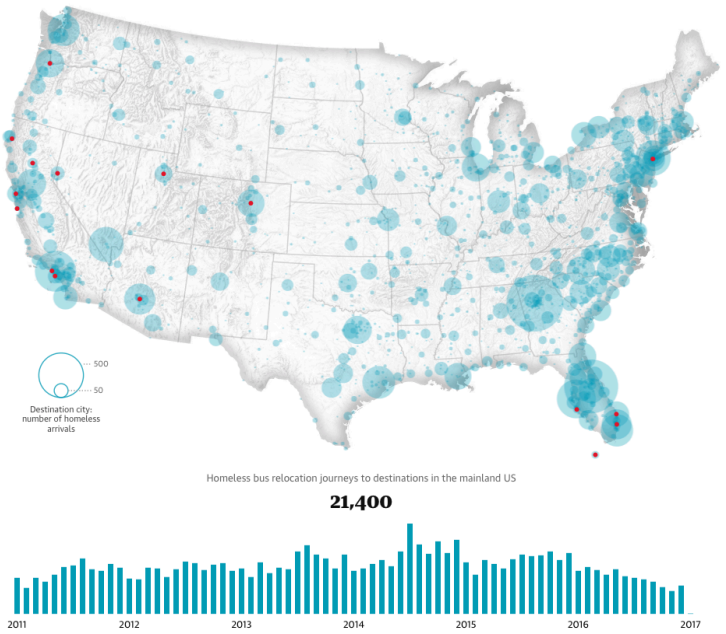


Figure 8. Screenshot of animation in the "Bussed out: How America moves its homeless story".[32]

Neither Bremer nor Wu are journalists but their expertise was crucial in bringing this 18-month project to life. They created the visualisation for the story but also worked with the reporters at a much earlier stage in the project, collating and cleaning the data and playing a crucial role in testing and understanding the data, thereby helping to create the backbone of the report.

> "We figured out the main story that we wanted to tell. And that was usually in the sense that they had a hypothesis and I would check it against the data and that might be true but it will have to be more nuanced. That's how it went back and forth, the entire day. By the end of the day we had like four subsections in one big big section of like the main story from like this," Bremer said. (Cairo, & Rogers, 2021, May 14)

This project was Wu's first time working with a media organisation and she encourages her clients to trust in the process: "Give us the freedom of creativity because you respect us for our expertise" (Cairo & Rogers, 2021, May 14).

However, if editors do not have a full appreciation of the field and how collaboration can enhance everybody's work, it can be challenging to persuade the hierarchy to play ball. Léchenet said that while working at France 2, the idea of collaboration was not appealing.

> "My boss was annoyed that we were not only sharing our information, but also that we were dependent on another media, and that we had to decide on a common publication

[32] Figure 8: Outside in America team, Bremer, N., & Wu, S. (2017, December 20). *Homeless bus relocation journeys to destinations in the mainland US* [Screenshot]. *The Guardian*. https://www.theguardian.com/us-news/ng-interactive/2017/dec/20/bussed-out-america-moves-homeless-people-country-study

> date, and that we no longer had control over publications." (Léchenet, 2021, MediaNumeric interview)

5.8.3. Domain Experts

It is important also to team up with people who are specialists in the field that is being researched. Without expert knowledge, it can be far too easy to miss the important points in the data, or the inconsistencies and errors. Shirley Wu, for example, says that there are certain topics that she will not touch by herself: "Some topics are so serious that I won't work on it unless I can partner with a domain expert because I know that I'm not qualified to do something just by myself" (Wu, 2021, MediaNumeric interview).

This is a view echoed in newsrooms everywhere where data journalists and visualisation designers team up with domain experts in areas of health, environment or finance, for example.

5.9. Who Should Learn Data-Driven Journalism?

Data is such an integral part of all areas of life today that everyone should learn basic data literacy regardless of whether or not they work in journalism and storytelling. Ideally, these classes should start in school, as described by the head of the Polish National Film Archive (FINA) Digital Repository, Joanna Kaliszewska in the following comments:

> "Statistics should be a compulsory subject at every university, but also the basics of programming, which should be part of every mode of teaching from primary school all the way to universities. I think it is also important to teach thinking in terms of cause-and-effect, which means, of course, logic, looking for connections and contexts." (Kaliszewska, 2021, MediaNumeric interview)

In addition to programming, Kaliszewska also sees the significance of addressing visualisations and the reading of information from a young age:

> *"I believe that we should teach data visualisation to children at a very early stage in pre-school education. In fact, functioning with information and interpreting information is a key skill, especially in the 21st century, when we're flooded by information and data from all sides. Finding one's place in this information overload is a very important competence."*
>
> - *Joanna Kaliszewska, Head of the FINA Digital Repository (MediaNumeric interview)*

Alberto Cairo says that in the Mediterranean tradition of schooling, such as in Spain or in Latin America, students are asked in middle school to choose between two pathways, the humanities or science subjects: "That's completely absurd, it's damaging to children," Cairo says (2021, MediaNumeric interview).

All the people interviewed for this report stressed that you only need maths taught at the level of a 12-year-old and an understanding of statistics in order to get started in the field of storytelling with data. All journalists today are faced with data sets, be they financial reports or government statistics, climate data or information on health or disease. It is not absolutely necessary to learn how to code or create data visualisations but it is vital that journalists are able to analyse data sets so that they can independently verify the information that is being provided to them:

> "You need to know how to calculate a percentage change, you need to know the difference between a median, a mean and a mode and when to use them. … Not knowing about those things is the equivalent of not knowing how to write grammatically." (Cairo, 2021, MediaNumeric interview)

Cook says that numbers are now just part of the job:

> "Reporters traffic in numbers whether they like it or not, just monetary figures and percentages, budgets, statistics, sports stats. It's impossible to avoid nowadays, especially as third parties and public relations officials and politicians are increasingly trying to use data on their side of things to nudge conclusions in a specific way." (Cook, 2021, MediaNumeric interview)

Léchenet says that working with data is actually nothing new. He cites as an example, *L'Express* which produced rankings of hospitals long before the discipline was called data journalism. He also cites *The Guardian*'s Datablog[33] created by Simon Rogers in 2009 as a real impetus for development.

It is clear from the research for this report that news and section editors and media bosses would also benefit from a better understanding of data-driven journalism if they have no direct experience of the field themselves. On joining *Libération* newspaper in 2015, Léchenet said "there was a lot of ambition and a lack of resources, and then there was, I think, a lack of understanding on the part of the editors of what we could do on the question of data and new formats" (Léchenet, 2021, MediaNumeric interview).

This view was echoed by Maxime Vaudano, who believes that this situation continues even today in France: "The lack of precision in the minds of managers as to what these new practices really are, and above all what they allow them to do at a journalistic level, leads them to be reluctant to invest." (Vaudano, 2021, MediaNumeric interview).

[33] Rogers, S. (2009, June 22). *Welcome to the Datablog*. *The Guardian*. https://www.theguardian.com/news/datablog/2009/mar/10/blogpost1

Karen Bastien, co-founder of the WeDoData data design studio concurs:

> *"I think there is a fear of engaging in this field because there is a lack of data culture among editors-in-chief. They don't really understand what it is — they only see the time-consuming aspect. It requires work from a lot of people, it's very expensive, and they ignore the editorial aspect that could never have been produced without all that work."*
>
> - *Karen Bastien, WeDoData, Co-founder (MediaNumeric interview)*

As detailed earlier in the report, analytics show high audience engagement with data-driven stories in some countries, which could enable media groups to monetise content. Even though data-driven stories can have wide-reaching impact and promote transparency and a fairer system for all, in many organisations the data team is still considered a niche area, separate from the core of the newsroom (Cairo, 2021; Léchenet, 2021; Vaudano, 2021, MediaNumeric interviews) and in France, the situation is even more acute.

> "I find that there is a distortion between the situation I observe at the international level – with turmoil in terms of data projects, an increasingly marked convergence between investigation and data, a large number of specialised freelance journalists, large editorial offices that invest resources and produce incredible things, etc. – and the French context, where major press managers aren't even remotely considering data projects." (Vaudano, 2021, MediaNumeric interview)

Working on data sets can take time, sometimes many months or even years. The International Consortium of Investigative Journalists worked on the Pandora Papers for two years before they began publishing stories. With budgets under constant pressure, editors who do not have a full understanding of the kind of impact that data can offer, do not know the tools or understand the process may find it hard to commit human resources to an expansion in data journalism.

> "I think that there is also perhaps a generational issue and that until we have a generation at the head of newspapers that 'got their hands dirty' with data in order to really understand what it is and to master it from an intellectual and practical point of view, data will remain this sort of 'black box' that scares people." (Bastien, 2021, MediaNumeric interview).

Léchenet provides an example of how unfamiliarity with data on the part of those in charge can create misunderstanding and frustration in the newsroom.

> "My boss asked me to estimate the speaking time of men and women in the National Assembly (French parliament), and to see if it had changed before and after the election. The database exists, and it's structured (nosdeputes.fr[34]). The formula is very simple, so in a day or two I had results and we could publish something. For my boss, this suggested that all work could be as quick, except that, for example, if we're going to do something on low-income housing finances, there's no accessible database. So either I ask for it, and it takes me six months, or we build it and it takes 3 weeks. In any case, it's not feasible in a day in the same way. And once we have the database, what does it tell us? It doesn't allow us to have a result in one day, because we are dealing with questions that are more complicated than: 'Is it a man or a woman speaking?'" (Léchenet, 2021, MediaNumeric interview)

However, media groups that have brought data into the heart of their news operations, where journalists have skilled up and returned to their departments with new tools, their reporting takes on an extra dimension.

"Data journalism is still seen as a niche area, as something that can be done or should be done just by a cadre of specialists within newsrooms," says Alberto Cairo, a pioneer in the field of data visualisation who led the Spanish newspaper *El Mundo*'s online team in the early 2000s when only a handful of publications around the world were pushing the boundaries of graphical work

> "I think that data visualisation is akin to writing. So, if you can learn how to write correctly, you can also learn how to design information graphics, visual explanations, data visualisations correctly. And some people would say: 'Well yes, but I don't want to be a visual designer, I want to be a writer.' And I say, if you can be a writer and also a designer and use both languages, then your work will improve significantly." (Cairo, 2021, MediaNumeric interview)

Simon Rogers, former news and data editor at *The Guardian* and Twitter, now data editor at Google News Lab, concurs.

> *"Data journalism is about using numbers to tell the best story possible. It is not about maths, or drawing charts or even writing code. It is about telling stories first and foremost – the maths and the charts and the code are all in service to that.*

[34] RegardsCitoyens. (n.d.). *NosDéputés*. NosDéputés. https://www.nosdeputes.fr/

You're no longer thinking solely about words. Instead this is about the best possible way to tell that story."

- Simon Rogers, Data Editor, Google (Rogers, 2014).

Other data and visualisation experts often say they were, or are still regarded as a "service desk," there to provide data or a visual in support of a text story rather than a source of ideas and an investigative team in and of itself. Léchenet said when he joined *Le Monde* newspaper in 2012, "the computer graphics department was in more of a supporting role and less about producing information" (Léchenet, 2021, MediaNumeric interview).

Craig Silvermann, now a reporter for ProPublica but previously a media editor of BuzzFeed News where he pioneered coverage of digital disinformation and media manipulation, says this deprives newsroom of a vital resource.

> "I think that one of the bad trends in newsrooms has often been to treat data journalists as a service desk where it's like: 'Hey, data friend, go pull this for me and then let me go do the reporting.' We don't treat Jeremy (eds: Jeremy Singer-Vine, data editor for the Buzzfeed News Investigative Unit) and his team like that. They are 100% full collaborators, and they just bring a whole other skill set and mind set." (Kelly, 2020, No. 1)

Without even developing data collection tools, coding or visualisation techniques, all journalists should have enough skills to feel comfortable and confident in challenging figures and statistics that are presented to them by outside sources on a daily basis. The quality of reporting would improve and it would foster transparency and help to rebuild public trust in the media.

5.10. Teaching Data Journalism

5.10.1. Challenges

The biggest issue that is systematically cited as a hurdle to learning data storytelling is a **fear of or a lack of interest in maths**. Most people involved in journalism or storytelling follow an education pathway based in the humanities. Teachers and instructors have to overcome a significant blockage in the acceptance of maths and statistics. "The biggest blockage that I run into is just the fear of math, the lack of confidence. I'll open a spreadsheet and they'll freeze, they look like deer in headlights," said Lindsey Cook (Cook, 2021, MediaNumeric interview).

Léchenet, who has been teaching data journalism for at least a decade, concurs:

> "I have the impression that people who come to journalism often have literary profiles which have been built up with the hatred, loathing, or at least with the lack of understanding of mathematics, and consequently we have people who, when they see a spreadsheet, say 'Ah, no, that's not for me. I don't understand it, I don't want to do that.'" (Léchenet, 2021, MediaNumeric interview)

5.10.2. Overcoming Hurdles

Instructors such as Cook and Lam Thuy Vo, who teaches at the Craig Newmark CUNY Graduate School of Journalism, recommend making the initial approach personal. Cook recommends asking students to create a spreadsheet with their contacts or sources, or to datafy an aspect of their daily life. Vo gets her students to download and then analyse data from their social media platforms.

Léchenet also takes a personal route with students, asking them to work with data related to their daily lives or a subject that they find particularly interesting in order to make it playful, and "above all without prejudicing the final result, i.e. not to use data to confirm a preconceived notion." (Léchenet, 2021, MediaNumeric interview)

Cook takes a similar approach to normalise spreadsheets by inputting your daily life into the documents.

> "One thing that I advise people to do is to integrate data and spreadsheets and numbers more into their daily lives to increase comfort with them, (…) using a spreadsheet to track how many people of colour you're interviewing, what percentage of your sources are women, what percentage are men. Just anything like that to increase confidence so that it's not such a scary thing to open a spreadsheet with 3,000 rows of data when you need to." (Cook, 2021, MediaNumeric interview)

Cooks says that **the bare minimum of knowledge to start training a journalist in working with data is a knowledge of Google Drive and Google Docs, basic computer skills, a bare minimum of maths literacy, such as fractions and percentages but a curious mind is crucial.** She stresses that the work is "just logic, at the end of the day."

She recommends going slowly to build confidence and says that if you teach too much information and too many tools in a short amount of time, the training does not "stick." She starts with the basics of spreadsheets, the four corners data check, what are columns versus rows, so that people become comfortable navigating spreadsheets. She also has students work with a partner during the training because peer learning is powerful for maths and computer science confidence.

Cook also recommends keeping examples, spreadsheets and lessons centred around journalism and the stories that reporters would be investigating in their day-to-day work lives: "It helps people realise that this is nothing but storytelling."

When Eva Constantaras starts a new training programme, she spends considerable time exploring her students' specialist subjects, working on topics, digging down into angles that would benefit from data. This approach gives people confidence, working in areas with which they are already familiar by simply introducing the new element of organising content into spreadsheets, ready for analysis (Kelly, 2021, Episode 32; Cairo & Rogers, 2021, Sept 28).

Rogers, who teaches at Medill in San Francisco also warns against focussing too much on the tools or the databases themselves. The key is in the mindset and understanding the potential.

> *"The tools keep changing but the skill sets, in a way, don't, like the things that you need to do, the attitudes towards the data, understanding what's possible."*
>
> - *Simon Rogers, Data Editor, Google (Kelly, 2020, No. 3)*

5.11. Where to Find Data Sets

The internet is awash with data sets. Most governments, national statistical institutes, public institutions, non-governmental organisations, companies and specialists make data sets publicly available. There are 250,000 data sets on the US government open data page data.gov[35] alone (Cook, 2019). Eurostat[36] claims to host 300 million statistical data about Europe on its interface. Sometimes a data set can be created by an ordinary person passionate about a specific topic. There are 13 different data sets, for example, about traffic signals on data.world.[37] Where to find data all depends on what you are looking for.

If the source of the data is not immediately obvious, the first port of call is GitHub[38] where most users put data sets that they have either collated or taken from a source and cleaned up. "If data exists out in the open, from my experience, it's quite likely that it's on Github," says John Burn-Murdoch, senior visual journalist at the *Financial Times* (Cairo & Rogers, 2021, June 8).

[35] U.S. General Services Administration. (n.d.). *The home of the U.S. Government's open data*. Data.Gov. https://www.data.gov/
[36] European Commission. (n.d.). *Eurostat*. Eurostat. https://ec.europa.eu/eurostat
[37] *There are 13 traffic signals datasets available on data.world.* (n.d.). Data.World. https://data.world/datasets/traffic-signals
[38] *Data Packaged Core Datasets*. (n.d.). GitHub. https://github.com/datasets

Buzzfeed, for example, makes all of its open-source data, analysis, libraries, tools, and guides freely available on GitHub.[39] They provide information about how the data set was compiled, the methodology and tips on how to read the set. Any user can explore the data they have collated and use it to find other stories, or to test stories that Buzzfeed has itself written about. The Sigma Awards, which celebrate the best in data journalism around the world, makes available on Github a database of all of the winners, citations and short-listed projects. Many other groups also share content on the platform.

There are also other websites that collate data sets. Shirley Wu, for example, always searches the Kaggle database.[40] Targeted search criteria in a web browser can usually reveal a goldmine of information as ordinary users make available information that they find fascinating.

It is however vital that journalists choose sources that are reliable. You should treat data sets as you would any other source of information, checking the credentials, the authority to speak about a certain subject. To ensure that a data set is reliable, it is vital to know the methodology that was used to create the data set. If necessary, it should be possible to retrace the steps that were taken to create the data set so that you can independently ensure the data is accurate.

Even if the data set is provided by a reliable institution, it is important not to skip the step of verifying the information yourself:

> "Sometimes you want to dig a little bit more because even if a good reliable source is frequently used by a lot of media companies, for example, there are always some parts that can be put in doubt. … You can have a global database that's, for example, working on a specific area, or a specific country, maybe the data is not the same level compared to other countries. So, it's really a job of trying to understand how the database has been set up and which are the original sources and what kind of a gathering reconciliation process has been done," said Alain Bommenel, head of infographics and data at Agence France -Presse (AFP). (Bommenel, 2021, MediaNumeric interview)

5.12. What Happens When Data Is Not Available?

While data is increasingly available in regions such as Europe, North America or Oceania, many journalists working in the Global South are handicapped by a lack of available data or by data that is considered unreliable. Journalists all over the world faced exactly this problem with the coronavirus pandemic that began in 2020. The pandemic was a turning moment for data journalism. With everyone searching for information and answers in the face of so much uncertainty, data journalism really came to the fore in explaining what was going on. Every single

[39] BuzzFeedNews. (n.d.). *GitHub - BuzzFeedNews/everything: An index of all our open-source data, analysis, libraries, tools, and guides.* GitHub. https://github.com/BuzzFeedNews/everything
[40] *Find Open Datasets and Machine Learning Projects | Kaggle*. (n.d.). Kaggle. https://www.kaggle.com/datasets

media organisation published charts and graphs on the daily basis showing, among many others, the daily infection rate, the death rate, the number of vaccinations and cases, etc.

However, when reporting on the pandemic during the first few months, journalists found that the data they needed was not available. Health services and public bodies were scrambling to treat patients. Few were working to aggregate reliable data, so other actors had to step into the breach.

Three teenagers in Melbourne, Australia – Jack, Wesley and Darcy – set up the Australian Covid tracking website, CovidBaseAU,[41] collating statistics from government sources and global data on infections, hospitalisations, deaths and vaccinations and organising them in an easy-to-read format.

The Agence France-Presse news agency used its reporters stationed all over the world to source local and national data on Covid cases and vaccinations. The data was verified and entered manually into their database.

The COVID Tracking Project[42] was one of the biggest data operations created to fill the gap left by holes in official data. *The Atlantic* journalists Alexis Madrigal and Robinson Meyer co-founded the project, setting a tracking spreadsheet when they realised that there was no national database in the United States that recorded how many people were being tested for Covid-19. They knew that the testing statistic was vital to understand how the virus was spreading throughout the country and which areas would be impacted next. They initially thought that the project would take a few weeks but more than one year later, the COVID Tracking Project had been the point of reference for media and US public institutions, including the government.

At the height of the project, there were approximately 500 volunteers and some 30 staffers scraping data from 56 different jurisdictions, often manually (Cairo & Rogers, 2021, The Data Journalist Podcast, No. 3), stitching together information from individual states in order to build up a national overview. They had to develop extensive processes[43] to manage the patchworked state reporting. They automated processes to validate and augment the manual work but stressed that if they had "set up a fully automated data capture system in March 2020, it would have failed within days."[44] While the manual data-entry was very labour intensive, the team says that this human process allowed them to truly understand the data in a way that would not have been possible with an automated collection.

---

[41] See: J., D., & W. (2021). *CovidBaseAU | About*. CovidBaseAU. https://covidbaseau.com/about/
[42] The Atlantic. (2021, March 7). *The COVID Tracking Project*. The COVID Tracking Project. https://covidtracking.com/
[43] Hoffman, H. (2021, April 28). *Analysis & updates | How We Entered COVID-19 Testing and Outcomes Data Every Day for a Year*. The COVID Tracking Project. https://covidtracking.com/analysis-updates/how-we-entered-covid-19-testing-outcomes-data/
[44] Gilmour, J. (2021, May 28). *Analysis & updates | 20,000 Hours of Data Entry: Why We Didn't Automate Our Data Collection*. The COVID Tracking Project. https://covidtracking.com/analysis-updates/why-we-didnt-automate-our-data-collection/

Gradually over time, the US federal government built its own database, benchmarking its results against those of the COVID Tracking Project. When the figures matched consistently and the journalists could prove that the figures were accurate, they decided to close down the project and the website. The need to maintain the database was over.

> "One of the reasons I think the project had so much impact was we both identified a need and delivered and shone the light on the new media versus traditional journalism but then it also filled a gap, not perfectly and we can talk a lot about the flaws but it filled it and what that meant was it put us in rooms and it put us in conversation with the people who were actually solving these problems, and I think that's really quite unusual," Madrigal said. (Cairo & Rogers, 2021, The Data Journalist Podcast, No. 3)

5.13. Interview the Data. How to See the Story in the Data?

Transparency is a core element of storytelling with data. In order to be sure that you are working from a database that is accurate, it is vital to submit the database through a series of questions and tests to evaluate the quality and accuracy of the information.

There are several ways of approaching a data set. Most journalists start with the topic or the question that they would like to investigate. From there, they go off in search of the information that can illuminate that story: background, experts, people to interview, the human story and the data. They use the data to explore the story, to test their theories, to prove or disprove. "The really kind of delicious data moments are where you get literally spreadsheets with numbers on it. And you interrogate that in a way that creates something that can illuminate an issue. That's when it gets really powerful." Peter Rippon, the editor of the BBC Online Archive, said (Rippon, 2021, MediaNumeric interview). However, this is not the only method.

Lindsey Cook of the *New York Times* says that there are many ways to approach data and it has a lot to do with how comfortable journalists are with data manipulation and analysis (Cook, 2021, MediaNumerics interview). In describing the most common process used by journalists who have less experience with data, Cook said:

> "For people who have a beat, it's often easiest to go the other way, so to start with a story that they want to tell or start with a question that they have. […] Perhaps they even start with interviews in a more traditional reporting sense and then they eventually come to the data."

Be curious, Cook continued:

> "I would consider myself more of a nerd, so I think that for me, it's just kind of fun to open a spreadsheet and to look for patterns: What's increasing? What's decreasing? What are the disparities? What are the outliers?"

> "There's definitely a lot of questions in the data and potential stories in the data that you can find, just by doing simple things like sorting, filtering and pivot tables. Those are where I start people and just with those three things, you can do a hell of a lot of data reporting."

Everyone stresses that the data is not the whole story: "The story is very rarely completely contained in the data," Cook said (2021, MediaNumerics interview). The data can direct you to the question and leads to an area that requires more investigation.

5.13.1. How to See the Story in the Data

Countless times, experts in the field say that the data is not the story. The story does not suddenly spring out of the data by itself. The story is in the data. But what story?

Lindsey Cook shows all of her students a video of a presentation by the Senior Product Manager of Tableau Public at Tableau Software Ben Jones. In the presentation,[45] Jones lists Seven Data Story Types which he describes as follows:

- Change over time
- Drill down
- Zoom out
- Contrast
- Intersections
- Factors
- Outliers

With one simple data set about freedom of the press in countries around the world and how that has changed over time, Jones shows how to create seven different stories, while adding that there are probably many other story types that could be extracted from the data. Among the examples Jones cites are the following:

1. **Change over time**: How has press freedom improved or worsened over time in the world, or in a specific region or a specific country? Why did those changes occur? What does that tell us about the societies in which we live?
2. **Drill down**: Focus on one specific country. Why did those changes happen in this one place?
3. **Zoom out:** Take the broad view. Look at the situation across the world or across a region, what can you infer from regional or global data?
4. **Contrast**: Compare the countries with the highest score of press freedom versus the lowest score. Where are the top 10 countries? Where are the lowest 10 countries? Where are they located, what does this say about those countries?

[45] See the video here: *Tapestry 2015 Short Stories - Ben Jones: "Seven Data Story Types."* (2015, March 13). [Video]. YouTube. https://www.youtube.com/watch?v=sEZj-eUfbNo&feature=youtu.be

5. **Intersections**: Where does data overlap? Where does the data cross over each together? Where does the data go from "less than something" to "more than something"? These are traditionally dramatic moments of change. What happened that caused this change?
6. **Factors**: Factors are things that work together to a higher level of effect. Components or pieces of a puzzle that combine together. What elements combined together can create a situation?
7. **Outliers**: Outliers are data points that sit outside the average range. They indicate that something different is happening with these data points. In Jones' example, which countries sit outside the average? Why do they sit outside the average? What makes them different?

"We really are fascinated by things that are different," Jones says in the presentation.

So from one simple set of data, Jones extrapolates many different stories and many different angles. This is, in part, why collaboration is so powerful in data-driven storytelling. Each person may see a different story in the data as they are coming to the information with their own vision, their own personal experience, culture and history. Editors should not be scared of collaboration, scared of competitors "stealing" a story. By collaborating together, it enables many more stories to be told and to provide a richer tapestry of the world.

5.14. Visualisation, Turning Data into Stories

Extensive studies examine how humans process and retain information. These are undertaken for a myriad of different purposes, for pure psychology or sociology, how best to persuade people or how to develop better sales practices. There are numerous works on how humans perceive colours and shapes. All these investigations feed into how to create effective data visualisations.

In an article[46] published on the website of the Poynter Institute for Media Studies, the Assistant Professor of Journalism at Hofstra University Russell Chun summarises six lessons for creating data visualisations gleaned from academic research.

The most striking finding for the purpose of data journalism specifically is a paper by researchers at Cornell University who found that simply including a graph in an article significantly increases reader persuasion. In their study, focussed specifically on persuasion in science, 68 percent of participants believed a scientific claim without a graph compared to a massive 97 percent of participants believed that same claim when a graph was included.

[46] Chun, R. (2015, October 27). *6 lessons academic research tells us about making data visualizations*. Poynter. https://www.poynter.org/reporting-editing/2015/6-lessons-academic-research-tells-us-about-making-data-visualizations/

The article also summarises research by William Cleveland and Robert McGill whose 1984 study into graphical perception[47] classified graphical encodings as follows, from most accurate to least:

1. position (dot plots, scatter plots)
2. length (bar and column charts)
3. angle (pie charts)
4. area (bubble charts)
5. colour (choropleth maps)

5.14.1. Presentation & Perception

How data is presented and visualised can transform the way a reader receives and understands a story. A long list of numbers in text format can be difficult to process and truly comprehend. Finding a visual way to represent the story can be the key to a reader swiping on to the next story or making them stop, sit up and take notice. All visualisation journalists have said that the Covid-19 pandemic was a turning point in the role of data in the newsroom. Graphics were vital in order to make sense of the sheer volume of numbers that emerged from the pandemic (Cairo. & Rogers 2021, Episode 3).

In 2012, the Environmental Defense Fund partnered with the communications agency Carbon Visuals[48] (now Real World Visuals[49]) to illustrate the greenhouse gas emissions produced by New York City. The three-minute animation[50] shows New York City gradually being buried under a mountain of blue spheres, each representing one ton of carbon dioxide.

[47] Cleveland, W., & McGill, R. (1984, September). *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods on JSTOR*. Jstor. https://www.jstor.org/stable/2288400 https://doi.org/10.2307/2288400
[48] Carbon Visuals. (2016, March 23). *Carbon Visuals*. http://www.carbonvisuals.com/all
[49] Real World Visuals. (2021, December 15). *Real World Visuals Projects*. https://www.realworldvisuals.com/projects
[50] Real World Visuals. (2012, October 19). *New York City's greenhouse gas emissions as one-ton spheres of carbon dioxide gas* [Video]. YouTube. https://www.youtube.com/watch?v=DtqSIplGXOA

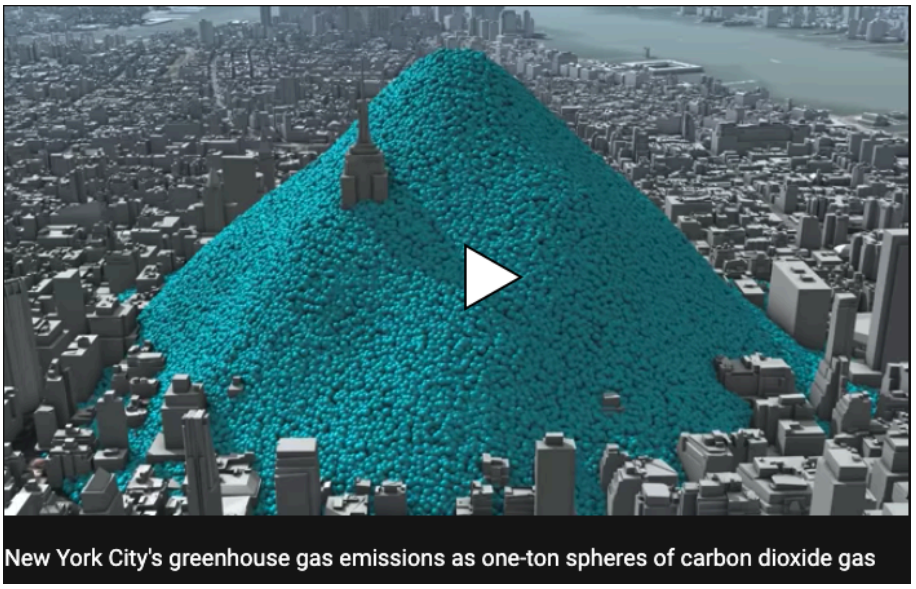


Figure 9. Image and clickable link to video illustrating New York City's greenhouse gas emissions. Source: Real World Visuals.[51]

The visualisation racked up more than 400,000 views on Real World Visuals' YouTube channel and the story was reported by media groups around the world, such as *The Guardian,*[52] NPR[53] and *Le Monde.*[54]

In an interview with the architecture and design magazine *Dezeen* in 2021, the agency explained the thinking behind the animation. "Carbon emissions are invisible and that's a core part of the problem," said Real World Visuals co-founder Antony Turner, "[...] if carbon dioxide was purple, we would start taking notice." (Hahn, 2021).

"Part of the problem is that some people are very cut off from quantitative information," said the agency's creative director Adam Nieman (Hahn, 2021), "[...] you put numbers and graphs in front of people and they bounce straight off. Our [aim] is to make the cause of climate change visible because very few other people are approaching it like that."

Time after time, using big data and creative visualisation, journalists are able to craft high impact stories which can bring about real-world change.

[51] Figure 9: Real World Visuals. (2012, October 19). *New York City's greenhouse gas emissions as one-ton spheres of carbon dioxide gas* [Video]. YouTube. https://www.youtube.com/watch?v=DtqSIplGXOA
[52] Rogers, S. (2017, July 15). *New York's carbon emissions visualised - as giant spheres*. *The Guardian*. https://www.theguardian.com/news/datablog/2012/oct/25/carbon-emissions-new-york
[53] Krulwich, R. (2012, November 17). *The Big Apple's Mayor Makes A Very Scary Video*. Npr. https://www.npr.org/sections/krulwich/2012/11/17/165275215/the-big-apples-mayor-makes-a-very-scary-video
[54] Foucart, S. (2013, March 13). *Manhattan noyée sous une pyramide géante de Co2*. Le Monde.fr. https://www.lemonde.fr/planete/article/2013/03/12/manhattan-noye-sous-une-pyramide-geante-de-co2_1846599_3244.html

5.14.2. Visualisation Rules

The myriad of studies and books have resulted in a certain set of rules about what makes a good data visualisation. One of the founding books in this field is *The Visual Display of Quantitative Information* by Edward Tufte (Tufte, 2001, Graphics Press). Tufte is an American statistician and professor emeritus of political science, statistics, and computer science at Yale University. He is widely considered to be one of the pioneers of data visualisation.

While it is important to know the rules, Alberto Cairo, who has himself written numerous books that are references in the field of data visualisation, warns about adhering dogmatically to rules which can be unnecessarily restrictive. He equates the creation of data visualisation to writing. You must learn the rules of writing, the grammar, its symbols and codes so that you can choose how to use them within your own form of expression.

> "For many years, visualisation has been taught in the past as a set of rules and these derive from some of the foundational books in data visualisation, such as Edward Tufte's *The Visual Display of Quantitative Information* and many others. Many people derive rules from those books that are completely undeserved, for example, don't use pie charts. Well sure, don't use pie charts if a pie chart is not appropriate but there are certain circumstances in which a pie chart might be appropriate.
>
> If the pie chart is the most intuitive way to represent a specific piece of information and if people understand it, then the graphic is good. It is not good prior to the designing of the graphic. A graphic is not good because of a theoretical rule. That's Platonism in reasoning." (Cairo, 2021, MediaNumeric interview)

There is a general consensus that pie charts are not a good form of visualisation. In the article *The Issue with Pie Chart: Bad by definition,*[55] Data to Viz explains that humans find it hard to read angles and when the values are similar, a pie chart is difficult to decipher correctly. The article demonstrates this point with three different data sets, represented in a form of pie charts, compared to bar charts. With the pie charts, it is difficult to perceive which colour has the largest slice of the pie. There is no such problem with bar charts.

[55] Holtz, Y. (2018). *The issue with pie chart*. Data to Viz. https://www.data-to-viz.com/caveat/pie.html

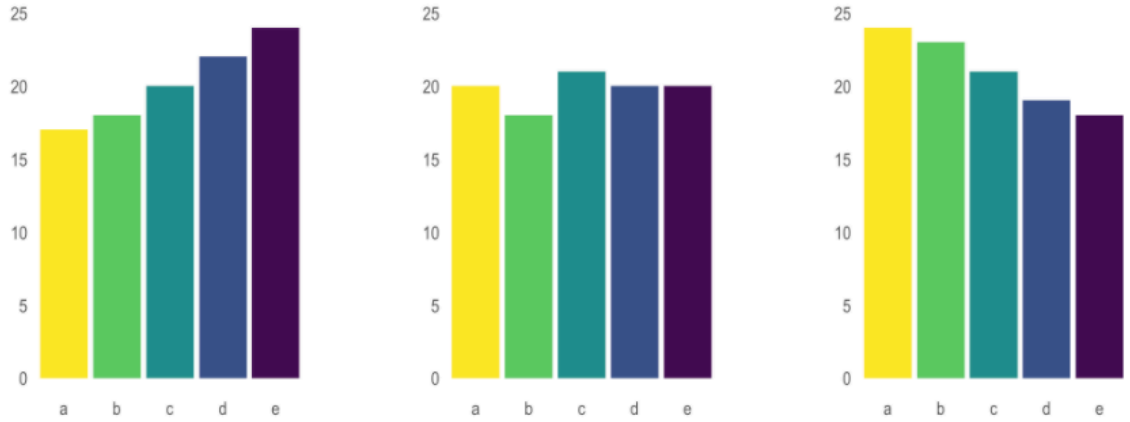


Figure 10. Comparison of the same information represented as pie charts and as bar plots. Source: Data to Viz.[56]

5.14.3. Be Honest, Be Ethical

It is entirely possible to create a data visualisation that is accurate, using correct data but which still misleads or gives a false impression (Malfatto, 2021, MediaNumeric interview). The decision in how to represent data is similar to the process facing a photographer, a cameraman, or a text journalist. Where a photographer or a cameraman chooses to place the camera (close to the ground, high in the air tilting down, a wide shot or a close-up) can totally change the impression of the image. By closing in on a group of people, it is possible to make a rally with only a small number of protesters appear as though it is a much bigger demonstration. Who a text reporter decides to interview, the figures they choose to include in their story and specifically what or who they choose to leave out can change the slant of an article. The same is true of data visualisation. The information you choose to include or exclude, the way you choose to present it, the colours you use, can give totally different impressions.

> "Some people devise bad visual models on purpose, to mislead their audience, but more often a faulty model is the result of a well-intentioned designer not paying proper attention to the data." (Cairo, 2016, Chapter 3. The Truth Continuum)

For example, the CEO of Apple Tim Cook was widely lampooned when he used a chart illustrating iPhone sales during a special Apple event on September 10, 2013.[57] About 20 minutes into the presentation, Cook projected a chart that he said showed how the latest iPhone model had taken their mobile phone sales to new heights.

---

[56] Figure 10: Data to Viz. (2018). *Comparison of barplots* [Graph]. Data to Viz. https://www.data-to-viz.com/caveat/pie.html

[57] *Apple Special Event. September 10, 2013.* (2013, October 9). [Video]. YouTube. https://www.youtube.com/watch?v=yBX-KpMoxYk

Figure 11. Chart of cumulative iPhone sales displayed during an Apple Special Event, Source: Apple[58]

At first glance, the chart seems impressive. However, on closer inspection, you see that there is no y axis with numbers to provide a scale with the number of devices sold. What is the scale that was used? It would be possible to generate an equally impressive curve moving from sales of just five phones in 2008 to 30 in 2013. To add an equally serious problem, the title indicates that the chart shows cumulative iPhone sales, not real iPhone sales per quarter. A chart for cumulative sales can only go up, as it is not possible to "unsell" phones that have already been sold. And as time goes by the curve will continue to chart significantly higher. The chart is an accurate depiction of accurate figures, but the choice the designer made in which figures to represent could, at best, be considered disingenuous, at worst misleading.

Quartz journalist David Yanofsky used Apple's quarterly iPhone sales reports filed with the US Securities and Exchange Commission and overlaid the results onto Apple's chart.[59] The figures look considerably less impressive.

[58] Figure 11: *Apple Special Event. September 10, 2013.* (2013, October 9). [Video]. YouTube. https://www.youtube.com/watch?v=yBX-KpMoxYk

[59] Yanofsky, D. (2020, June 24). *The chart Tim Cook doesn't want you to see*. Quartz. https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

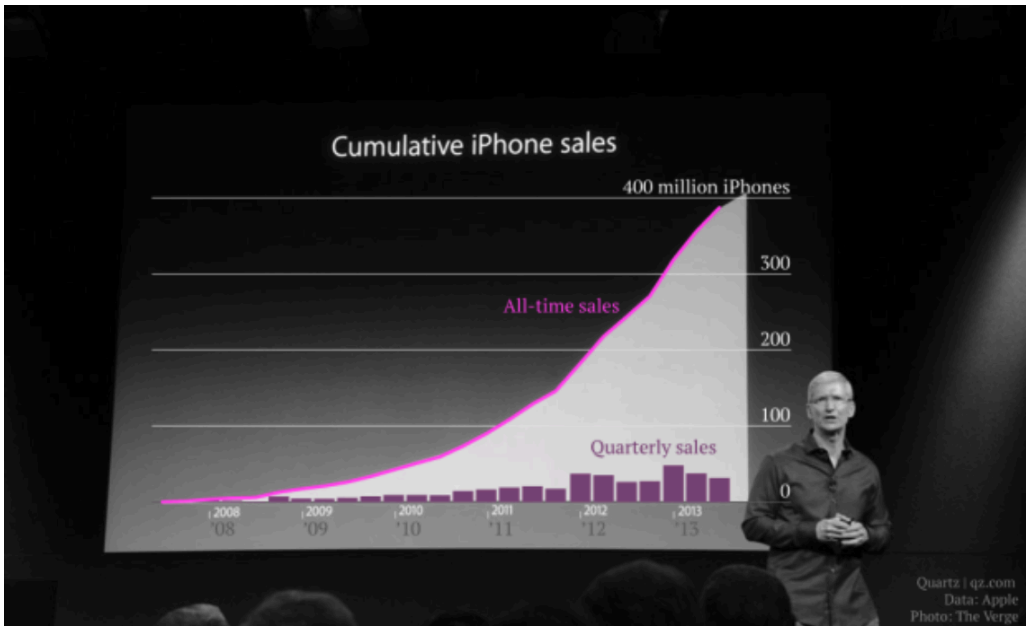


Figure 12. Quarterly iPhone sales are overlaid with the cumulative iPhone sales: Source: David Yanofsky.[60]

There are countless examples of visualisations being manipulated to serve a message rather than the reality. From tilting charts to making them 3D to give a false impression of size and ratio, to modifying the y or x-axis to disproportionately change the scale, from not starting the y-axis at zero to skipping over time periods for which there is no data or creating a break in a bar chart.

Both Alberto Cairo and Daniel Levitin cite in their books (Cairo, 2019, Introduction; Levithin, 2018, pp. 29-30) a graph that Fox News used in 2012 to illustrate how much more tax the wealthy would have to pay if the George W. Bush-era tax cuts to the top rate of federal tax were allowed to expire.

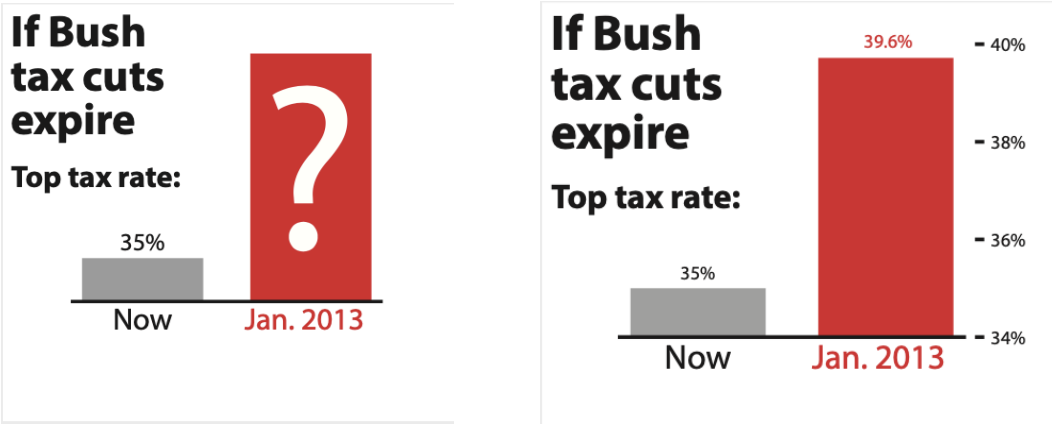


Figure 13. Bar chart used by Fox News in 2012 that illustrates the tax increase wealthy people will pay if Bush-era tax cuts are allowed to expire. On the left, the bar chart with no scale. On the right, the same bar chart shows a scale of just five percentage points. Source: Courtesy of Alberto Cairo.[61]

[60] Figure 12: Yanofsky, D. (2020, June 24). *The chart Tim Cook doesn't want you to see*. Quartz. https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

[61] Figure 13: Cairo, A. (2020, January 8). *All graphics from "How Charts Lie" freely available in two color schemes*. The Functional Art. http://www.thefunctionalart.com/2020/01/all-graphics-from-how-charts-lie-freely.html

In the bar chart, Fox News chose to start the scale at 34 percent rather than 0, which is the recommended practice. The scale ends at 40 percent rather than the usual 100, which means that the difference in the scale is just five percent. As the scale is abnormally shortened, it distorts the difference between the two figures and the increase is excessively exaggerated.

Below is the bar chart with the scale set at zero. The difference between the two columns is much smaller and the tax increase does not appear as significant.

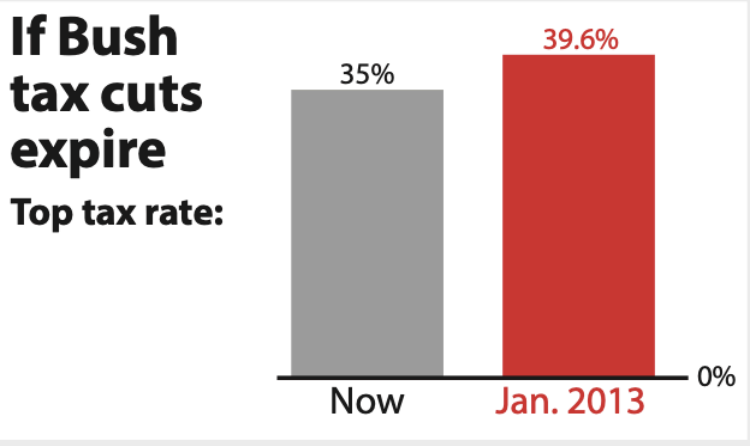


Figure 14. Bar chart illustrating possible tax increases with the scale set to zero. Source: Courtesy of Alberto Cairo.[62]

5.14.4. Societal Codes

Many of the visualisation rules stem from the way that humans perceive colour, shapes and forms. Red for example has a very long wavelength and so is one of the most visible colours on the colour spectrum. It is therefore used to grab attention and signal a warning or danger. It is the colour of blood so when we see it, it can mean injury. It can be associated with heat (fire), rage, shame (blushing), attraction and passion (flushed face). As humans, we are genetically programmed to pay attention to red and data visualisation can use this to an advantage, drawing attention to the most important part of the visualisation.

Figure 15. Stock images of red traffic signs.

[62] Figure 14: Cairo, A. (2020, January 8). *All graphics from "How Charts Lie" freely available in two color schemes*. The Functional Art. http://www.thefunctionalart.com/2020/01/all-graphics-from-how-charts-lie-freely.html

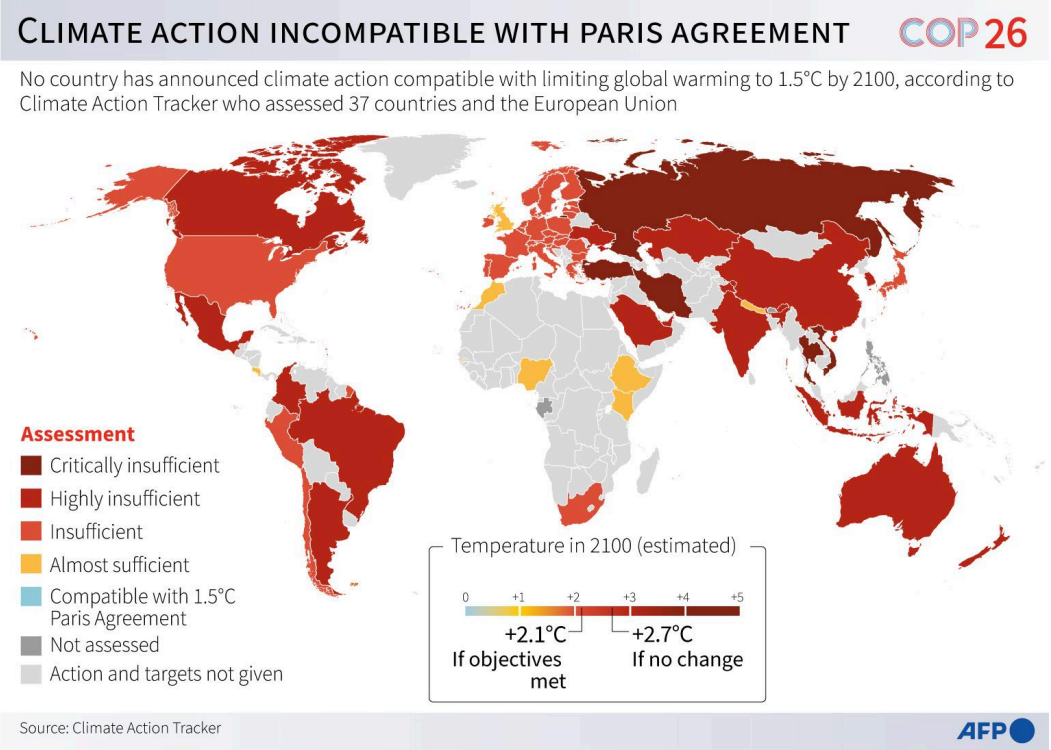


Figure 16. Map showing climate action announced by 37 countries and the European Union following the COP26 climate summit, according to Climate Action Tracker. Source: AFP/Valentina Breschi, Kun Tian.[63]

But not all cultures perceive colours in the same way. For example, in western countries white is often associated with purity and innocence (christening, confirmation and bridal gowns), cleanliness and sterility, (the medical profession's white coats), it is the colour of peace (the white dove or the white flag of surrender). Generally, the colour has positive associations. In eastern countries however, white is often worn at funerals and during the mourning rites and can be associated .

[63] Figure 16: Map showing climate action announced by 37 countries and the European Union following the COP26 climate summit, according to Climate Action Tracker. Source: AFP/Valentina Breschi, Kun Tian

Figure 17. (On left) Britain's Catherine, Duchess of Cambridge holds Britain's Prince Louis of Cambridge on their arrival for his christening service at the Chapel Royal, St James's Palace, London on July 9, 2018. Source: POOL/AFP/Dominic Lipinski[64]

Figure 18. (On right) Dressed in traditional Korean mourning white, the sisters of Park Mi-Jin, one of scores of young salesgirls killed in the collapse of the Sampoong Department Store, help their grieving mother (C) to the funeral on 3 July near Seoul's Kangnam Hospital. Source: AFP/KIM JAE-HWAN.[65]

In the left image above, the Duchess of Cambridge is dressed in white to celebrate the christening of her child Prince Louis (Source: Dominique Lipinski, AFP/POOL). The photo on the right shows a family grieving during the funeral of a young sales girl killed in the collapse of a department store in Seoul. They are dressed in traditional Korean mourning white (Source: Kim Jae-Hwan, AFP).

How colours are used in data visualisation can tap into very ancient and fundamental aspects of human nature and culture and designers must take this into account.

5.14.5. Direction of Travel

Many scripts of the world are read from left to right. A timeline is also read from left to right. We instinctively believe that up indicates an increase (and that is usually good) while down indicates a decrease (which is usually bad). Yet for scripts that read from right to left, the opposite is true.

[64] Figure 17: (On left) Britain's Catherine, Duchess of Cambridge holds Britain's Prince Louis of Cambridge on their arrival for his christening service at the Chapel Royal, St James's Palace, London on July 9, 2018. Source: POOL/AFP/Dominic Lipinski

[65] Figure 18: (On right) Dressed in traditional Korean mourning white, the sisters of Park Mi-Jin, one of scores of young salesgirls killed in the collapse of the Sampoong Department Store, help their grieving mother (C) to the funeral 03 July near Seoul's Kangnam Hospital. More than 200 are still missing beneath the rubble of the store. Source: AFP / KIM JAE-HWAN

A quick glance at the graphics below, which track annual fossil carbon dioxide emissions and projections, a reader could believe that they show two opposing situations. On the left, fossil fuel CO2 emissions are increasing (which in this case is not good at all) while on the right, they could seem to be falling. In reality, both graphs show exactly the same data, with the same scale, the same timeline and the same axes. The only difference is the graphic on the left is in Latin script which is read from left to right while the graphic on the right is in Arabic, which is read from right to left. It shows the importance of studying a visualisation rather than making an inference from a brief glance.

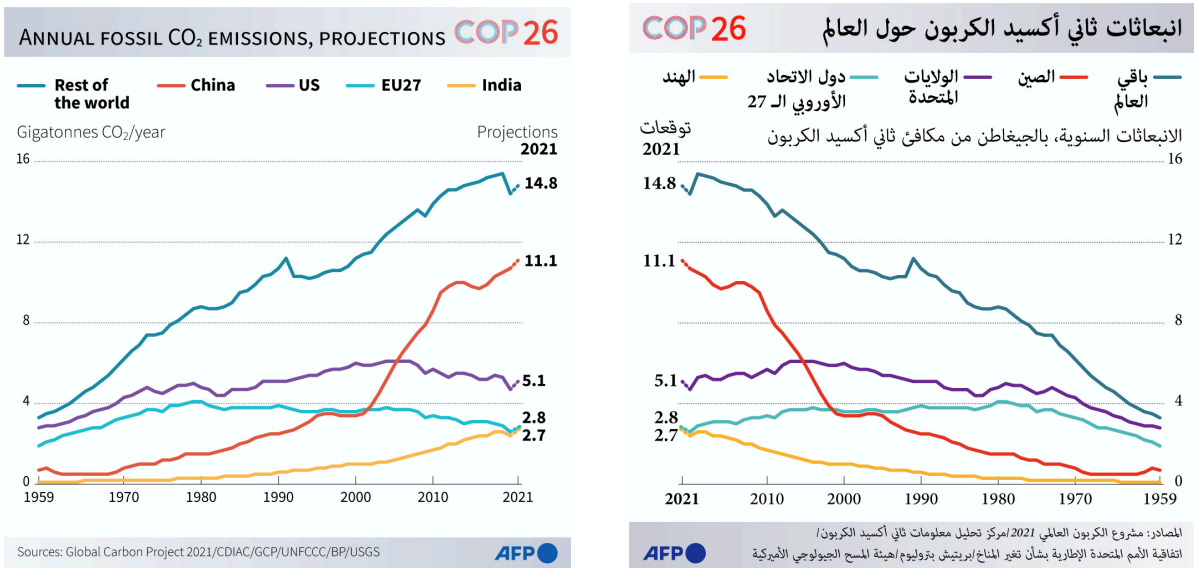


Figure 19. Chart showing annual fossil carbon dioxide emissions and 2021 projections, according to the Global Carbon Project 2021. Source: AFP/Valentina Breschi, Gal Roma.[66]

When creating a visualisation, it is vital to keep in mind for whom you are creating the graphic, your audience's reference point and the cultural conventions in interpreting visualisations.

It is not possible within this report to detail the endless ways graphics can mislead. In his most recent book, *How Charts Lie: Getting Smarter about Visual Information* (2019), Alberto Cairo delves into these practices in depth so that everyone can learn to become more graphicate.

There are also many websites that call out bad practices in visualisation. Venngage provides a succinct but effective run-down,[67] there is a Wikipedia page[68] devoted to misleading graphs and on reddit, there is a thread called Data Is Ugly[69] where users point out misleading visualisations. There

[66] Figure 19: Chart showing annual fossil carbon dioxide emissions and 2021 projections, according to the Global Carbon Project 2021. Source: AFP/Valentina Breschi, Gal Roma

[67] McCready, R. (2021, November 17). *5 Ways Writers Use Misleading Graphs To Manipulate You [INFOGRAPHIC]*. Venngage. https://venngage.com/blog/misleading-graphs/

[68] Wikipedia contributors. (2021, December 1). *Misleading graph*. Wikipedia. https://en.wikipedia.org/wiki/Misleading_graph

[69] Reddit contributors. (2012, November 23). *Data Is Ugly*. Reddit. https://www.reddit.com/r/dataisugly/

is much to learn from these books, websites and threads about the pitfalls in this field and how to improve your own practice.

5.16. Tools

There are a huge number of different types of software that can be used in storytelling with data. Some will be used for data collection, others for organising and analysing data and more still to create visualisations. There are tools designed to be used by people with little or no programming or coding skills, others that are for people with more advanced skills and knowledge. The following tools are the most common software used but it is by no means an exhaustive list.

The first tools are to organise data. Most people use spreadsheets, such as Microsoft Excel or Google Sheets. The COVID Tracking Project also used Airtable. Excel has the ability to handle massive data sets with hundreds of thousands of entries but Sheets offers greater flexibility in terms of collaboration and its ability to interface with non-Google design software such as Flourish or Datawrapper. A spreadsheet is the most important tool to learn how to use proficiently. It is important to understand how to organise, filter and analyse the data, how to use pivot tables and identify the anomalies, disparities, differences and outliers in the data. Spreadsheets and pivot tables will allow you to find the story in the data, to test your theories about a particular topic.

If the data is not readily available, you might have to scrape it from a database. There are many different tools to do this but the most common among data journalists R or Python. Python is considered one of the easiest programming languages for a beginner to learn. People more familiar with coding may use a programme called Node.js, a JavaScript runtime environment that works outside of the browser and is used primarily for writing server-side code.

The data may not even be conveniently placed in a database or spreadsheet. It may be buried in hundreds of pages of text pdfs so the International Consortium of Investigative Journalists developed Datashare, a free open-source desktop application which allows journalists to simultaneously search pdfs, images, texts, slides or any other files. It can also automatically detect and filter by people, organisations and locations. The ICIJ also secures documents from third-party interference. Pdftotex is an open source toolkit created by Xpdf which can also be used to extract text from pdf files.

There are a whole range of different programmes for data visualisations. Datawrapper, Flourish and Tableau enable people with no coding or programming skills to create charts, maps and tables simply uploading data from a spreadsheet. For more control, others use the R project for statistical computing, KNIME, JavaScript or D3.js which helps bring data to life using HTML, SVG (Scalable Vector Graphics), and CSS (Cascading Style Sheets).

There are also tools to automate tasks, such as Zapier used by the COVID Tracking Project.

There are many tools that are free to use and tutorials are widely and freely available online in order to learn how to use them.

Tools most frequently cited by data journalists and the people interviewed by the MediaNumeric consortium are:

- Microsoft Excel
- Google Sheets
- Python
- Datawrapper
- Flourish
- R
- Linkurious
- JavaScript
- Tableau
- D3
- Datashare
- Pinpoint

Although not directly related to working with and creating stories from data, Slack is another important tool to adopt. Slack is an instant messaging communications platform where users can send a message to a specific person, persons or to a whole group. The chat app is an extremely effective way to communicate and share files. Discussions can be grouped into channels and specific workspace. There are many Slack channels dedicated to data journalism. It is an important and precious resource for everything that is going on in the field. Some of the channels are specialised on specific topics, but others look at the field from a broader perspective. The Slack channel of the Sigma Awards,[70] which celebrates the best in data journalism across the world, is free to join and is a place of lively discussion for everything that relates to the field.

Some have also cited project management as an important skill to develop (Wu, 2021, MediaNumeric interview). When working with large data sets, it is easy to become overwhelmed by the amount of information to process. It is useful to learn how to divide up the roles in an investigation, assign tasks and develop a methodology to approach the complex stories.

5.17. Challenges for the Future, Archiving and Diversity

Alexandre Léchenet cites the lack of available positions and advancement as an obstacle to a wider uptake of data journalism in the newsroom.

> "The media have very few positions to offer in terms of data, and these positions are still rather 'all-purpose'. It is difficult to have fun by doing visualisation on the internet, investigation, and a little advanced analysis all at once. This frustration sometimes drives

[70] *The Sigmas Slack team*. (2021, January 19). The Sigma Awards. https://sigmaawards.org/slack

> young journalists to change jobs. In fact, there is more training than there are jobs." (Léchenet, 2021, MediaNumeric interview)

As data journalism is still considered a niche area, training is not widespread. It can be hard for overworked newsrooms to carve out the time for journalists to learn these new skills. The 2022 Data Journalism survey found that more than one in three of the 1,800 respondents were entirely self-taught (European Journalism Centre, 2023).[71]

Beyond training incoming journalists and skilling up existing newsrooms, one of the main challenges for the future of data storytelling is archiving. There are countless projects that have been lost because the tools that were used to create the work have been retired. Adobe Flash or Google Fusion tables are just two such examples of tools that no longer exist.

All data journalists and storytellers who have been working for at least a decade say that some of their work no longer exists because the technology on which it was built has been retired, the media organisation folded, the URL link no longer works, the rights over website domain name expired or the website changed, underwent an upgrade or the project was not considered a priority in allocating server space. "It has reinforced to all of us how fragile our work is, that things that I worked on eight years ago just aren't there anymore. They don't exist," Rogers said (Kelly, 2020, No. 3).

This is a major problem in terms of history of the investigation, the data sets and transparency of the workflow.

The COVID Tracking Project was such a mammoth, collaborative project (see section 5.11) about such an extremely significant moment in world history that its creators felt that it was vital the project was saved for posterity. The COVID Tracking Project ended when federal agencies were able to provide the correct Covid-19 figures for the United States themselves. In order to ensure that they would be able to always access the project, they are putting the project into flat html. For this project, they even archived the Slack workspace and channels so that in the future people will be able to track how the team came to the decisions that they made:

> "We are able to show a lot of the context for; How is this decision made? Why were Texas' hospitalisation numbers frozen for five days in July? Well, we can show you. Here's why and here's the connection, you can actually see the conversations that occurred in Slack about those things," Madrigal (Cairo & Rogers, 2021, The Data Journalism Podcast, Episode 4).

Many other projects will not benefit from this level of dedication to archiving. It is a point of order to try to find a way to save this work for the decades to come.

[71] European Journalism Centre (2023) *The State of Data Journalism 2022*. European Journalism Centre. https://datajournalism.com/survey/2022.

Open-source data is another major challenge. Although most people recognise the benefit to democracy to have open source data, many companies are now realising the monetary value of their information. X, the platform previously known as Twitter, now charges a fee for access to its data. Data is also running up against General Data Production Regulations (GDPR). In November 2022, the European Court of Justice ruled that the public access to the list of people registered as beneficiaries of commercial companies registered in Luxembourg contravened the right to privacy and Luxembourg closed it down (Poujol, V., 2022, Nov. 24)[72]. Without access to this database, the money laundering and tax evasion at the heart of Luxembourg's financial centre would never have been revealed.

[72] Poujol, V., (2022, November 24) *Coup dur pour la transparence financière*, Reporter. https://www.reporter.lu/luxembourg-cour-de-justice-ue-coup-dur-pour-la-transparence-financiere/

6. Fact-Checking: The Information Ecosystem

6.1. Main Issues in Information Ethics

Rumours and fabrications have existed since the dawn of time. In periods of confusion or fear, conspiracy theories have always claimed to offer explanations to momentous events. The ancient Roman politician Octavian's (Augustus) well-documented rumours about the intentions and slurs on the character of his rival Mark Antony in 33BC show that political propaganda is perhaps as old as politics itself. The media have always played a role in the transmission of exaggeration, hoaxes, and falsehoods. In 1835, *The New York Sun* printed a series of fantastical news stories describing life on the moon that shocked their readers. The Great Moon Hoax is an early example of a fabricated news story that dramatically increased the newspaper's circulation. "False information, misperceptions and conspiracy theories are general features of human society" (Nyhan 2020, p. 232).

Despite its long history, in 2016 so-called "fake news" and its potential impact on the democratic process became an issue of global concern. On November 16, 2016, Craig Silverman at BuzzFeed News reported that the top 20 most popular "fake news" stories had gained more engagement on Facebook than the top 20 real news stories during the final three months of the US presidential campaign. The fake stories totalled 8,711,000 engagements compared to 7,367,000 for the real stories. The most popular made-up story, one that suggested that the pope had endorsed the candidacy of Donald Trump, gained almost a million (960,000) engagements on the platform (Silverman 2016).

Fabrication circulating virtually was provoking real-world action. A month after the election on December 4, 2016, a young man from North Carolina walked into a pizza restaurant in Washington DC armed with an AR-15 rifle and fired three shots. Thankfully, no one was hurt. Following his arrest, the man, Edgar Maddison Welch, told police that he was there to investigate for himself claims related to the "Pizzagate" theory. Pizzagate was a baseless conspiracy theory that circulated online, amplified via social media that falsely claimed that senior Democratic politicians, including Hillary Clinton, were running a child sex ring from the restaurant (Lipton, 2016).

The information profile of the 2016 US presidential election along with the other momentous election campaign of that year, the Vote Leave Brexit campaign in the United Kingdom, brought online information pollution to the attention of governments, organisations and eventually to the platforms themselves.

The term "fake news" was chosen as the Collins Dictionary word of the year for 2017, signalling the prominence of the issue in the public consciousness. However, people studying the field quickly found that "fake news" is little more than a vague term used to describe everything from satire and parody, to false balance in the media, to manipulation and propaganda. Wardle and

Derakhshan (2018) therefore argue that it is a woefully inadequate description to capture the complexity and diversity of information disorder.

Worse still the term "fake news" has been appropriated by politicians, like Donald Trump, and the right-wing media that support them as means of discrediting legitimate stories that they find unwelcome. As Habgood-Coote (2018) writes: "In the mouths of right-wing demagogues, the accusation is a command not to believe a story and to distrust the institution that produced it." Indeed, exposure to elite discourse on "fake news" promotes distrust of the media and makes real news content harder to identify (Van Duyn and Collier 2019). In co-opting the term to discredit unwelcome news stories right-wing actors are undermining the very democratic value and commitment to the truth that they profess to uphold.

The need for terminology that captures the complexity of the information disorder has therefore been recognised. As Baybars Orsek, Director of the International Fact-Checking Network, explains "The terms are constantly shifting. In 2016 it was all 'fake news' and then it became all misinformation and now everybody prefers to call it disinformation. It's basically a bad information problem" (Orsek, 2021, MediaNumeric interview).

In "Fake News. It's Complicated" Claire Wardle of First Draft News describes seven different forms of misinformation and disinformation that cover a diversity of problematic information (Wardle, 2017):

- Satire or parody - no intention to harm but can mislead or confuse
- Misleading content - misleading use of information
- Imposter content - content that seeks to impersonate a genuine source of information
- Fabricated content - content that is totally false and designed to deceive and harm
- False connection - headlines, visuals or captions don't support the content
- False context - genuine information placed in a false context
- Manipulated content - genuine content that is manipulated to deceive

Following Wardle and Derakhshan (2018) these forms of problematic content and more, fall within three broad categories that are important to distinguish:

- Misinformation - information that is false but not intended to cause harm
- Disinformation - information that is false and deliberately created to cause harm
- Malinformation - Information that is genuine and used to inflict harm

Although an important feature of the information landscape, a discussion of malinformation, like revenge porn, information leaks or doxxing (the act of releasing previously private personal information about someone into the public domain with an intent to harm), is beyond the scope of this report. The key distinction is between misinformation and disinformation (Burger, 2021) and is a distinction that rests in the intention of the creator and those who share it. In some cases, it might be possible to establish the intention behind the creation of problematic content, but often it is a case of trying to discern the thoughts in people's heads. While it is important to recognise

the distinction between, for example, false health advice based on a misunderstanding of the underlying science versus deliberate attempts to convince people that proven science is wrong and should be ignored, fact-checkers rarely investigate the intentions behind the claims that they check. They are more concerned with verifying the claim. Therefore, in what follows, unless specifically talking about disinformation, we will use the terms misinformation, unverified, false or misleading claims, allowing that what we refer to as misinformation may actually be disinformation.

6.2. The Information Landscape

Fact-checkers overwhelmingly believe that the misinformation problem has become worse and that they face a greater volume of false and misleading claims now compared with 10 years ago. What was once a fringe phenomenon now feels like what Eugene Kiely, director of FactCheck.org, describes as a “firehose of misinformation” (Kiely, 2021, MediaNumeric interview), a view echoed by Angie Drobnic Holan, editor-in-chief at PolitiFact, who believes the primary issue is: ”There's just an extraordinary amount of misinformation in people’s environment that is difficult to avoid” (Holan, 2021, MediaNumeric interview).

It is extremely difficult to measure the scale of misinformation in the environment. Researcher and Africa Check founder, Peter Cunliffe-Jones, puts that down to the complexity of the ecosystem whereby firstly, a lot of misinformation spreads offline, peer-to-peer, and secondly information moves fluidly between offline and online spaces, across different social platforms, through the traditional media and then back into community networks in ways that are very difficult to track. “One of the problems for assessing whether or not the situation has become worse, which is the general assumption that it's become worse, is that we don't really know,” said Cunliffe-Jones (Cunliffe-Jones, 2021, MediaNumeric interview). There has so far been no systematic review that has looked at changes in the volume of misinformation over time.

However, the digital age has undoubtedly added layers of complexity and capacity to the information ecosystem. Equally, the way that information is created, shared, and consumed has changed radically.

The internet has removed traditional barriers to publishing making it vastly cheaper and easier. It has given a platform to everyone to say whatever they want, whenever they want. ”It's not just a few people in the centre speaking about things. It is coming from lots of different directions,” says Andy Dudfield, head of automated fact-checking at Full Fact (Dudfield, 2021, MediaNumeric interview). This means that a lot more information is being created and moves across different platforms and different venues in diverse ways.

“I think people are overwhelmed with information” (Holan, 2021, MediaNumeric interview) and this avalanche of information has laid bare people’s perceived lack of media literacy. ”There seems to be public confusion about how to separate misinformation from truthful facts” (Holan, 2021,

MediaNumeric interview). An Ipsos MORI poll published in March 2021[73] showed that 30 percent of Europeans reported they found it either somewhat or very difficult to distinguish between true and false content online. Fifty-eight percent of Europeans surveyed in the same poll declared an interest in learning about tools to distinguish between true and false information online (Archer, 2021).

Travelling across and through social media platforms, the reach of information has been supercharged as it seamlessly traverses the globe. It is shared effortlessly and can reproduce exponentially; one false claim can go viral in an instant. Technological developments can deliver that information to the palm of your hand. Not only can information be created and disseminated from anywhere and at any time, you can receive and share that information anywhere and at any time. Angie Drobnic Holan says people must now make a special effort to avoid misinformation in their environment (Holan, 2021, MediaNumeric interview).

In an examination of approximately 126,000 verified true and false news stories shared in 4.5 million tweets on Twitter between 2006 and 2017 Vosoughi et al. (2018) found "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news" (p. 1146). Or as the 17th–18th-century author Jonathan Swift put it: "Falsehood flies, and truth comes limping after it."

The impact of the increased creation and sharing of information is amplified by the algorithms that social media platforms use to curate and deliver content to their audiences. "Social media algorithms have amplified misinformation allowing it to travel much further considerably increasing the scale of misinformation and disinformation," says Sophie Nicholson, deputy head of the AFP Fact Check team (Nicholson, 2021, MediaNumeric interview).

The picture is complex. The current information ecosystem allows for more information to be created and shared by diverse actors across online and offline networks which in turn increases the reach and the impact of all types of misinformation, false and misleading content. In the following sections, we will take a more detailed look at the character of the misinformation, topics where misinformation is particularly rife and which venues create and share it. We'll review some of the factors that make people susceptible to believing misinformation and finally examine the harm it can cause.

6.3. Subjects Prone to Misinformation

Misinformation is spread over and through a wide variety of topics. False and misleading claims can be made about anything but through analysis broad themes do emerge. Peter Cunliffe-Jones and colleagues from fact-checking organisations across Africa recently built a database of fact-checks from a typical six-month period and analysed the content. They identified 20 topics

[73] Ipsos MORI survey on media literacy:
https://www.ipsos.com/ipsos-mori/en-uk/online-media-literacy-across-world-demand-training-going-unmet

representing four key themes (Cunliffe-Jones et al, 2021). The topics are highly contextual. The specific issues or incidents generating misinformation within broad themes are culturally specific and historically contingent. There is also considerable cross-over between the themes with topics subject to misinformation for more than one set of reasons.

6.3.1. Politics

The first generation fact-checking organisations focussed mainly on political speech. Much of the misinformation was connected to perennial political questions of governance, policy, crime or the state of the country, often from one side of the political debate about the other.

This can intensify during an election season and often extends to misinformation about specific political personalities or candidates, like the repeated misinformation regarding Barack Obama's country of birth and more recently about the current US vice president Kamala Harris.

Elections themselves are a source of misinformation with false claims frequently found about how or where to vote or coordinated disinformation campaigns from paid-for services or even official state actors, such as the well-documented Russian interference in election processes in the United States and Europe (Nicholson 2021, MediaNumeric interview).[74]

As Emmanuel Vincent from Science Feedback explains, false and misleading claims are often made about any subject that is polarising; a subject that is weaponised in some way and recruited to fight a political battle. Examples from a US context could be climate change, abortion or gun control; all subjects that have been appropriated as party political issues.

6.3.2. Emotional Topics

Topics of misinformation are often emotional subjects. Social issues that are weaponised to fight political battles inspire strong feelings (Kiely, 2021, MediaNumeric interview). These could again include abortion, which has a religious element to it, gun control, or debate around US second amendment rights.

One subject that interviewees felt was a particularly strong source was race and immigration; "those topics are continually subject to misinformation and in a very divisive way" (Holan 2021, MediaNumeric interview). Angie Drobnic Holan suggests that in the American context this is down to "anxieties about the changing racial makeup of the US, and those messages tend to play on white fears, to be more specific about it" (Holan, 2021, MediaNumeric interview). These messages are seen as a particular concern of the far-right political parties making immigration and race prominent campaign subjects. Peter Burger of Leiden University, founder of Dutch fact-checking organisation Nieuwscheckers, says in the Netherlands there are "… a number of radical right-wing parties in the parliament and quite a number of people on social media, on Twitter on Facebook,

[74] Also see Senate Intelligence Committee report on Russian interference in the 2016 U.S presidential election: Volume 1 https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf and Volume 2 https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf.

making unfounded claims about immigration" (Burger, 2021, MediaNumeric interview). Health claims, particularly related to unsupported so-called cures or nutritional advice for example, are another broad topic of misinformation that draw strength from anxieties or fears related to our own bodies (Vincent, 2021). Health misinformation is also important because of the direct impact it can have on people (Burger, 2021, MediaNumeric interview). Fears of violence, child abduction and strangers stoked by false rumours spread on WhatsApp combined in India in 2018 to spark vigilante action that allegedly resulted in the deaths of more than a dozen people (Safi, 2018).

6.3.3. Vulnerabilities that Attract Opportunism

Fears are a vulnerability that can be exploited. Vulnerability and need are seen as other themes that draw opportunistic misinformation. In African countries where there is a lot of youth unemployment, Peter Cunliffe-Jones and his colleagues have seen a lot of fake job offers phishing for personal details, like identity, phone or bank details, or fraudulently demanding a small fee (Cunliffe-Jones et al, 2021a). These types of criminal activities can be seen everywhere through email scams or fraudulent text messages designed to resemble official communication from banks. Another common type of scam exploiting needs are hoax claims to appeal for donations, perhaps related to accidents or emergencies (Cunliffe-Jones et al, 2021a).

6.3.4. Things of Concern in the Moment

Sophie Nicholson sees parallels with the news calendar (2021, MediaNumeric interview). National elections, for example, put hot-button political issues in the spotlight in a more intense way. Specific incidents like a mass school shooting in the US can resurrect or intensify false and misleading claims around gun control (Kiely, 2021, MediaNumeric interview) while the Covid-19 pandemic has generated endless false claims about vaccines (Holan, 2021, MediaNumeric interview).

Sometimes old claims, referred to as zombie claims, return in a modified form to connect to peak news events, for example, the myth that 5G technology, which had been subject to suspicion months before the emergence of the novel coronavirus, was in some way responsible for the coronavirus pandemic (Rahman, 2020a). Cell phones, and cellular technologies like 5G and its predecessors have been topics of related misinformation for years, linking them to various pathologies from swine flu to brain tumours. Accidents, emergencies, and crises in the news are all topics of the moment that will generate new claims or resurface old ones.

Angie Drobnic Holan has also observed that misinformation comes in what she describes as "weird trends" that seem random and gain short-lived attention like a recent slew of false claims about serial killings or reporting the death of celebrities when they are still alive (Holan, 2021, MediaNumeric interview).

Misinformation about the behaviour and views of popular celebrities also belong in this category of 'things in the moment'. The viral deep fakes of US actor Tom Cruise on TikTok were just one of

several recent high-profile examples where this type of technology was used to target public figures.

6.4. Where Does Misinformation Come From & Why?

Misinformation is created, shared and amplified by a wide variety of actors. These actors range from political elites and institutions, to the media, organisations with vested interests and the general public. Wardle and Derakhshan (2018) propose three categories of incentives to produce and distribute misinformation: political; financial and social and psychological with actors working across these categories. In this section we will present the main categories of actors who are producing and sharing misinformation and examine the incentives that might be driving it.

6.4.1. Political Elites & State Actors

While not all political actors spread misinformation political elites as a class of communicators are seen as one of the key progenitors of misinformation as they and their staff craft messages to progress their political agendas. Politicians and governments can do this through online and offline messaging and work through the media as well (Cunliffe-Jones, 2021, MediaNumeric interview).

The Washington Post Fact Checker team closed their count of Donald Trump's false and misleading public claims at more than 30,000 for his time in office. The former US president may be exceptional how voraciously he used misinformation in both public appearances and on social media in order to serve his own ends but he is by no means alone in this. "Though exceptions exist (for example, conspiracy theories about 9/11), [political] elites have played a key role in creating or popularising many of the most salient misperceptions of recent years." (Nyhan, 2020, p. 227).

While political actors like Donald Trump have been recorded repeatedly pushing totally unverified claims and absolute falsehoods, political misinformation can often be more subtle and more grounded in shades of fact and kernels of truth. Full Fact analysed data from their own experience of fact-checking UK political speech and made a report on the problematic way that statistics are used to support arguments and policy messaging. Statistics are used out of context so as to be misleading or can be cherry picked so as not to give a true picture of reality. During the coronavirus pandemic, for example, statistics were repeatedly quoted but without the source of the figures. These techniques are used to score political points; to claim credit for their own side or lay blame at the door of the opposition (Full Fact, 2020b).

Political actors can use their platform to spread misinformation personally, but they can also extend their reach by employing professional companies to do it for them (Cunliffe-Jones, 2021; Nicholson, 2021, MediaNumeric interviews).

The latest annual global survey of organised social media manipulation by the Oxford Internet Institute shows that governments, public relations firms and political parties are now producing misinformation in a way that is becoming more professionalised and on an industrial scale. The

report shows that the production of political misinformation is widespread; it was deployed as part of political communication in 76 out of 81 countries surveyed. It also showed that the role of strategic communications firms in spreading political misinformation is growing; state actors employing such services were detected in 48 countries. And, it is big business; state actors spent almost $60 million on firms who use bots and other amplification strategies to create the impression of trending political messaging. Cyber troops, government or political party actors who seek to manipulate public opinion online, have spent over $10 million on Facebook advertisements (Bradshaw, Bailey & Howard, 2021).

Social media manipulation and misinformation is not only used by state actors as part of their domestic strategies, these techniques are also used as a tool of geopolitical influence. The report notes that "In 2020, for example, authoritarian countries like Russia, China and Iran capitalised on coronavirus disinformation to amplify anti-democratic narratives designed to undermine trust in health officials and government administrators" (Bradshaw, Bailey & Howard, 2021. P. 2). Russia's efforts to manipulate democratic elections in the US for example has been well documented.[75]

6.4.2. The Media

Political misinformation has broad reach through social media channels but domestic media ecosystems play a key role in amplifying and conveying political disinformation. "The role of the mainstream media as agents in amplifying (intentionally or not) fabricated or misleading content is crucial to understanding information disorder" (Wardle and Derakhshan, 2018. P. 25).

In a discussion of the media we should draw a distinction between traditional mainstream media that may have political biases that guide editorial policies but that operate within basic norms of journalistic practice and a hyper-partisan right-wing media ecosystem that includes major networks and smaller alternative news media. Both create, spread and amplify misinformation and disinformation but in different ways.

Angie Drobnic Holan points toward the right-wing media, websites and talk radio as key players in the US disinformation ecosystem and polarising influence that allows misinformation to breathe. While they promote right-wing political themes there may also be a financial incentive to spreading misinformation that enrages and engages their audience (Kiely, 2021, MediaNumeric interview). A study by Yochai Benkler and colleagues (2020) from Harvard University suggests that Donald Trump's unsupported claim about election fraud was a highly effective disinformation campaign that was driven by elites through the mass media, with social media playing a secondary role. These findings echoed an earlier study by the authors on the political media ecosystem in the US between 2015–2018 where they found that the right-wing media with Fox News at its core was far more successful at spreading false beliefs than Russian trolls or clickbait sites promoted on social media (Benkler et al., 2020; Benkler, Faris and Roberts, 2018).

[75] See the Senate Intelligence Committee report on Russian interference in the 2016 U.S presidential election: Volume 1 https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf

The traditional media are also particularly highlighted as a source of health and science misinformation through a lack of knowledge and expertise. Peter Burger believes that within mainstream media journalism sloppiness and lack of training in how to understand and communicate science is a major problem:

> *"Quite a few of them tend to accept, for instance, press releases at face value; will just use one source and not dig into the original publication; lack the skills to judge statistics… journalists in general could do better."*
>
> - *Peter Burger, Associate Professor in Media Studies, University of Leiden (MediaNumeric interview)*

The term false balance refers to media reporting on issues like climate change, for example, that puts established scientific consensus on an equal footing with unverified claims by interest groups to show "both sides of the argument" (Scheufele and Krause, 2019). Amplifying minority non-factual arguments in the name of balance where no serious scientific debate exists could lead to public confusion; audiences could hear or read about alternative explanations or denials of phenomena, like climate change or an unverified link between vaccines and autism, for example, by commentators given a platform by the media for the sake of "balance" or "objective reporting" and conclude that the science is uncertain (Vincent, 2021; Boykoff & Boykoff, 2004; Dixon and Clarke, 2013). According to reports the BBC in 2018 issued new guidance and made training available to its journalists for reporting on climate change warning them to avoid false balance after the corporation was censured for failing to challenge climate sceptics in interviews (Carrington, 2018; 2017).

Headlines are the most prominent part of news stories. In the digital age when shared on social media the headline can become separated from the rest of the story which lies behind a link. If a user doesn't click through the link to read the whole story, the headline becomes the entire message. A study by researchers at the French National Institute for Research in Computer Science and Automation in collaboration with colleagues at Columbia University found that 59 percent of links shared on Twitter were never clicked through and remained "silent" (Gabielkov et al., 2016). If the headline of a news story forms the entire message over half the time, then it becomes vital that headlines are accurate (Full Fact, 2020b). Full Fact's study of the UK press regulator's accuracy rulings shows that this is not always the case. They discern five ways in which headlines were found to be misleading, exaggerated or in rare cases simply unsupported by the article itself (Full Fact, 2020b).

6.4.3. Clickbait Artists & Industries

Another vector of online misinformation are the misleading clickbait advertisements. These advertisements commonly found on news sites use shocking or false headlines or promises of

miracle cures to drive traffic. A recent example classified as clickbait by Polish fact checkers Demagog includes an incendiary false quote attributed to Jan Duda, father of Polish President Andrzej Duda (Demagog, 2021). This type of misinformation generates money through pay-per-click advertising arrangements. A systematic study of advertising content on mainstream news sites and misinformation sites classified 44.6 percent of the advertising content they sampled as problematic; containing deceptive, low-quality or misleading content and they found no significant difference between legitimate news sites and known misinformation sites in terms of the amounts of problematic advertising content they carried (Zeng, Kohno & Roesner, 2020).

Organisations with vested interests may be engaged in spreading misinformation for financial reasons (Emmanuel Vincent, 2021). The $4.5-trillion wellness industry, including supplements, lifestyle brands and alternative therapies is often accused of spreading false, unsupported and misleading health claims. For example, CNN reports that Gwyneth Paltrow's lifestyle brand, Goop, paid civil penalties of $145,000 following legal action from prosecutors in California over “unsubstantiated claims” relating to the company’s products (Ravitz, 2018). Similarly, it has been alleged that the petroleum industry engaged in a disinformation campaign to cast doubt on the science behind warnings over climate change; a campaign likened to that of the tobacco industry’s efforts to cast doubt on the science behind the link between cancer and smoking (Lawrence, Pegg & Evans ,2019).

6.4.4. General Public

While elite actors and the media may be responsible for driving the creation of misinformation the general public play a role in sharing and amplifying those messages (Grinberg et al., 2019). While misinformation sharing can easily occur offline within peer groups most studies of how the general public are sharing misinformation occurs with analysis of users behaviour or behaviour intention on social media platforms.

One of the principal reasons is seen as political divisions: “Affective polarisation is a major driver of spreading disinformation, spreading fake news on social media; the people hate their political opponents so much that they sling any negative news at them that they can get their hands on, be it real news or fake news” (Burger, 2021, MediaNumeric interview). A recent study of 2,300 US Twitter users came to the same conclusion highlighting the power of political partisan attitudes as the main motivation to share political fake news (Osmundsen et al., 2021).

In a series of experiments with social media users Gordon Pennycook and colleagues found that participants were well able to judge the veracity of news headlines and reported a strong preference for only sharing accurate information online yet the veracity of news headlines had little effect on participants’ sharing intentions. Investigating this apparent disconnect they found that using a prompt to prime the participant to consider the accuracy of a news headline significantly improved the quality of the information shared. The researchers suggest that people might share misinformation online because something about that environment distracts the user from considering accuracy (Pennycook et al., 2021).

Away from the polarising effects of politics and the wild claims of the clickbait artists, many people may share unsupported claims or inaccurate advice online simply because they believe them to be true. Misleading or false health advice may be shared on social media by users simply in an effort to be helpful (Cunliffe-Jones, 2021, MediaNumeric interview). A post on Facebook containing a list of accurate and inaccurate advice about Covid-19 was shared hundreds of thousands of times and migrated to other platforms like WhatsApp and Twitter. In one version the post said it was from a relative of the author who worked at a hospital. Full Fact, who fact-checked the post, noted that the author edited the post to remove some of the false claims and updated after the fact-check was published (Rahman, 2020b).

Social media provides the architecture for instant reinforcement and reward when sharing content as posts count likes and re-shares. The approval of the online community can be a motivator to sharing content that the sharer may even know is inaccurate but finds entertaining or confirms evidence of group identity (Cunliffe-Jones, 2021, MediaNumeric interview).

6.5. What Factors Make People Vulnerable to Misinformation?

Truth judgements are constructed from three types of information: base rates, consistency and feelings. Firstly, people have a base rate bias toward accepting new input since most information they encounter in their lives is true; it takes cognitive effort to disbelieve. Secondly, people judge the truth of new messages against what they already know; things that are consistent with facts in their memory are accepted more easily. Thirdly, people interpret their feelings as evidence of truth; if the message feels good or familiar and easy to process then it feels more true (Brashier & Marsh, 2020).

Each of these factors increases overall accuracy in general and are good strategies for interacting with the world. Yet people hold to banal misconceptions like “you can see the Great Wall of China from Space”; believe things that have been shown to be untrue like “Barack Obama was born outside the US”; disbelieve messages for which there is abundant evidence like “human activity is causing climate change”; and even accept claims that seem improbable like “the Covid-19 vaccine contains a microchip inserted by Bill Gates.”

People make mistakes. The ways in which people process information and make judgements lead toward predictable illusions and biases. These biases can sometimes be manipulated and illusions induced. Here we consider some of the factors that make people vulnerable to misinformation and some of the processes that can lead people to accept false claims and form misperceptions about the world.

Exposure to misinformation is the first step to believing it (Nyhan, 2020). The more you are exposed to a claim the truer it seems, even improbable ones. “A reliable way to make people believe in falsehoods is frequent repetition, because familiarity is not easily distinguished from truth” (Kahneman 2011, p. 62).

In a much cited study, Fazio et al. (2015) found that students were more likely to rate false claims like “a sari is the name of the short pleated skirt worn by Scots” or “the Atlantic ocean is the largest on Earth” as truer when they had been exposed to them before.

Even a single exposure to a claim is enough to boost its credibility. The more people are exposed to a claim the more familiar and comfortable it feels. Repetition increases the fluency of the claim making it easier to process. Inferring truth from a feeling of familiarity is an example of a heuristic, a type of cognitive shortcut to save brain power. This illusory truth effect through repetition has been demonstrated for all types of claims from trivia to fake headlines, lasts over time and is a potentially powerful route to false beliefs (Brashier & Marsh, 2020; Pennycook & Rand, 2021).

Psychologists have found that you don’t even need to repeat an entire claim or fact to make it appear true. For example, being repeatedly exposed to the phrase “the body temperature of a chicken” makes people more likely to accept the statement “the body temperature of a chicken is 144 degrees” (Kahneman, 2011).

Another important cue that we use when judging the truth of new information is the source (Pennycook & Rand, 2021). People are more likely to believe information from a trusted source. One study into this effect showed that identifying Donald Trump as the source of a false claim increased Trump supporters’ belief in that claim and decreased the belief of his opponents (Swire et al., 2017). The authors suggest that “people use political figures as a heuristic guide to what is true or false” (p. 1); trust in the source makes the message seem more true.

Seeking to avoid bad consequences is perfectly healthy. But, sometimes in making judgements, especially with regards to risk, people can be led astray by directly engaging their emotions without realising it; intuitively asking themselves ‘How do I feel about this?’ rather than ‘What do I think about this?’. The effect heuristic as it is known can lead people to overestimate the risk of a bad outcome based on their emotional response to it. For example, in a classic study Paul Slovic and colleagues found that people judge tornados as a bigger killer than asthma, death by lightning strike as more frequent than botulism and death by accident as more likely than stroke when health statistics show that the opposite is overwhelmingly true in all cases (Lichtenstein et al., 1978).

Warped estimates of causes of death are seen as caused by the availability of news in the media about tornados and deadly accidents and people’s strong emotional reactions to that news make them feel more dangerous (Kahneman, 2011). So when faced with a false claim that presents a emotionally charged bad outcome, for example “vaccines cause miscarriage,” people can intuitively perform a substitution and answer the question, “how do I feel about miscarriages?” rather than “what do I think about this claim.”

A recent study shows that people who reported heightened emotionality at the beginning of a task were more likely to believe false (but not true) news headlines and that instructing participants to rely on emotion increased their belief in false (but not true) news headlines. This second finding

suggests a causal link (Martel et al., 2020). This is important since as we have previously discussed a lot of misinformation is spread through subjects that evoke strong emotions.

People also seem vulnerable to misinformation that is consistent with their worldview or identity. When a position on societal questions becomes a symbol of group identity or political affiliation, like for example whether human activity is causing climate change, people may employ identity-protective cognition which leads them to find reasons to believe in content that is consistent with their partisan position and be sceptical of content that is inconsistent. This is a form of motivated reasoning which predicts that those with greater understanding will use it to find arguments to support their own political viewpoint.

To test this theory Dan Kahan and colleagues set participants in the US a difficult task drawing inferences from a set of data. Those with the highest level of numeracy did best when they were told that the data related to the efficacy of a skin cream, a non-partisan subject. But, when participants were told that the same data concerned the effect of gun-control on crime, a high profile and divisive issue, responses became politically polarised and less accurate. The tendency to skew responses toward the subject's political affiliation was greatest in the most numerate (Kahan et al., 2017).

Pennycook and Rand (2019) find evidence that people who engage in more analytical thinking as measured by the Cognitive Reflection Test (CRT) evaluated the truth of news headlines more accurately. Reminders to consider the accuracy of news headlines also have been found to reduce the sharing intention and real-world sharing on social media of false news (Pennycook et al., 2021). This seems to suggest that sometimes a lack of thought can make people susceptible to believing and sharing misinformation. The authors suggest attention-based interventions on social media platforms to counter the spread of online misinformation.

6.6. Consequences of Bad Information

It is commonly held that misinformation has the potential to cause harm on an individual and societal level. It can misinform our opinions, mislead our actions, cause threat to life, distort government policies and disenfranchise the vulnerable in a variety of ways. It is important to recognise that not all misinformation causes identifiable harm. Research suggests that up to 40 percent of misinformation shared is harmless (Cunliffe-Jones et al., 2021a). But that still suggests that more than half can cause harm and while direct causal links are difficult to establish, evidence suggests that potential and identifiable effects and associations are wide-ranging and complicated.

Peter Cunliffe-Jones and colleagues have studied the effects of misinformation and document potential and identifiable harms that they categorise into ten fields at different levels of society. At an individual level they found that misinformation can place physical and mental health at risk, can limit participation in democracy and can harm individual finances. At a more societal level it can

entrench negative stereotypes, cause harm to the natural world and even distort government policies (Cunliffe-Jones et al., 2021a).

In what follows we will take a more detailed look at the identifiable and potential harms in a few of those key areas.

6.6.1. Democracy

Political misinformation and disinformation is a key concern. Bradshaw et al. (2021), the authors of the 2020 global inventory of social media manipulation report mentioned earlier are unequivocal:

> *"The manipulation of public opinion over social media remains a critical threat to democracy."*
>
> *- Samantha Bradshaw, Hannah Bailey, Philip N. Howard (2021)*

The report demonstrates that the potential threat to disrupted democratic processes is on an upward trajectory; the practice is becoming more widespread, professionalised and more industrialised with each year. The proliferation of misinformation is "hugely dangerous and it's dangerous for democratic governance in some very specific ways" (Holan, 2021, MediaNumeric interview).

Since the 2016 election of Donald Trump and Britain's decision to leave the EU, and the alarm around online misinformation, researchers worried about the link between misinformation and specific political outcomes have produced mixed results. Some have argued that worries about direct effects on voting intention might be overstated (Cunliffe-Jones, 2021, MediaNumeric interview) especially since during the 2016 US presidential election false online stories made up only a small proportion of the electorate's election media diet (Allen et al. 2020). Others find evidence that belief in false news stories impacted vote choice in favour of right-wing populist parties for example during the 2017 German parliamentary election (Zimmermann & Kohring, 2020).

Misinformation has a long history of being deployed to disrupt electoral processes. The storming of the US capitol on January 6, 2021 in an attempt to prevent the certification of the results of the US presidential election by supporters of Donald Trump fuelled by the repeated unsupported claim that the election had been "stolen" and linked to the QAnon conspiracy theory is a recent and emblematic example.

False and misleading claims can also be used as an instrument of voter suppression. In the United States context, as elsewhere, this often targets disenfranchised communities, including minorities, low-income and immigrant communities (Vandewalker, 2020).

One common tactic is to circulate misinformation about where, when and how to vote. For example during the 2018 midterm elections in the US false voting information circulating on social media included instructions to vote by text, which is not valid anywhere in the US and claims that because of potential overcrowding at polling places voters of one party should vote the day after Election Day (Kim, 2018).

False messages can also attempt to dissuade people from attending polling stations entirely by falsely claiming that immigration officers are checking voters identities, there is violence around the polling place or a false claim on Twitter: "warning that everyone over age 60 that #coronavirus has been reported at ALL polling locations for #SuperTuesday" (Vandewalker, 2020; Cunliffe-Jones, 2021, MediaNumeric interview).

Misinformation in political speech and about specific policies can cause public confusion over what is true and what false, what Benkler, Faris and Roberts (2018) refer to as "disorientation." Research by Full Fact and BritainThinks in 2019 suggests that lack of trust in politicians is leading to voter disengagement. In their survey 58 percent of respondents said that they didn't trust what they hear from politicians and 19 percent said that misinformation had put them off voting entirely (Full Fact, 2020b).

A further worry is that misinformation is driving and reproducing political polarisation, where facts are less important than an entrenched worldview (Nicholson, 2021, MediaNumeric interview). According to Peter Burger speaking about the situation in the Netherlands "there is an increased polarisation that is tied to disinformation" (Burger, 2021, MediaNumeric interview).

At a more general level misinformation can distort meaningful political debate and distract from the deliberative activities of government. Angie Drobnic Holan explains that:

> "If you can't agree on the facts you can't come to any kind of agreement on the common good or solutions or practical next steps… you can't compromise either because you don't even agree on the basic facts. So, misinformation short circuits democratic governing processes in a very pernicious way that I don't think is fully appreciated." (Holan, 2021, MediaNumeric interview)

6.6.2. Public Health

Misinformation is not only disruptive to democracy but can potentially have a serious effect on public health.

Misinformation has always had the potential to undermine public health efforts. The discredited link between the MMR vaccine and autism that was proposed in a now retracted scientific paper in 1998 had a devastating effect on vaccination rates in the UK. Recent outbreaks of measles in the US, where it was once eradicated, and Europe, have been blamed on inadequate levels of vaccine take up.

Early in the global coronavirus pandemic the World Health Organisation (WHO) led warnings that governments must be prepared to combat the "infodemic"; an overabundance of information that contains false and misleading information about Covid-19 and the virus that causes it. The concern was that a global spread of unreliable or misleading information could pose a major threat to public health by causing confusion, mistrust of health authorities and therefore undermine governments' health response.

To examine the common themes in Covid-19 misinformation, fact-checking organisations from five European countries analysed 645 of their articles from the month of March-April 2020. They found that misleading advice about cures, unverified claims about the severity of the disease, unsupported advice on prevention behaviours, misinformation about vaccination and conspiracy theories linking the origin of the virus to Bill Gates or 5G cellular technology were all strong themes common to most or all countries.[76] If people follow health advice that is ineffective or fail to follow good advice through misperception or a mistrust of medical expertise then harm at an individual and societal level is likely to follow.

Researchers in Canada concluded that exposure to misinformation on social media increased misperceptions about the virus which reduced both adherence to disease prevention and control behaviours and support for life-saving policies (Bridgman et al., 2020). Others have linked misinformation about vaccination to false beliefs and lower trust in medical experts (Stecula et al., 2020).

Although falling across the world, vaccine hesitancy is still at high levels in some countries. An Ipsos MORI poll conducted in March 2021 showed that 41 percent of people surveyed in France reported that they did not intend to accept the Covid-19 vaccine (Lacey, 2021). The picture in other countries showed 11 percent in Brazil, 22 percent in Australia, 25 percent in Germany and 35 percent in the USA mistrustful of the vaccine. Studies using data from the US and the UK have shown that reluctance to accept the Covid-19 vaccine is associated with exposure to scientific-sounding misinformation about the virus and the disease it causes (Loomba et al., 2021).

When bad information distorts government health policy the effects can be disastrous. It is estimated that around 330,000 lives were lost between 2000 and 2005 because of the South African government's refusal, in the face of scientific consensus, to accept the link between HIV and AIDS and to implement a lifesaving treatment programme with antiretroviral (ARV) drugs which the government branded as poisonous (Chigwedere et al., 2008).

6.6.3. Physical Violence

Misinformation and propaganda are features of war and conflict, but false rumours spread online and offline have also been shown to lead to mob violence and vigilante killings in peacetime.

[76] See the report I*nfodemic Covid-19 in Europe: A Visual Analysis of Disinformation (2020),* put together by AFP, CORRECTIV, Pagella Politica/Facta, Maldita.es & Full Fact. https://covidinfodemiceurope.com

In October 2019 three Ethiopian scientists conducting research into intestinal worms and trachoma in a village called Gonji in the country's Amhara were attacked by the local community who believed a rumour that they were there to poison the village's children. Two of them were killed and the third only escaped with his life because the villagers thought he was dead (Nur, 2019). No one knows how the rumour started.

The scale of this type of harm can explode when the rumour spreads online. In 2018 India was wracked by dozens of mob lynchings of strangers blamed on false rumours of child kidnappings spread over WhatsApp. At least 30 killings over a one-year period all over the country were linked to an out-of-context video clip from a Pakistani child safety campaign (Safi, 2018).

6.6.4. Economic Harms

Just as misinformation about diseases and cures can harm your physical health false claims relating to businesses and products can harm companies' economic health and market value. Hoaxes and false rumours on social media targeting companies like Tesla and Starbucks and Wetherspoons forced companies to put in place communication strategies to quickly shut down rumours before they can do too much damage (Atkinson, 2019).

For example, in May 2019 Metro Bank's share price plunged 11 percent before they could quash false rumours circulating on WhatsApp that the company was about to or had gone bankrupt. The rumour sparked a run on the bank with customers queuing in branches to empty their safety deposit boxes (Makortoff et al., 2019; Brown, 2019).

In the United States, Dominion Voting Systems, the company at the centre of unsupported claims about vote rigging during the 2020 US presidential election filed separate lawsuits against Donald Trump's lawyers Rudolph Giuliani and Sidney Powell and Fox news seeking damages of over a $1 billion for each. The company contends that false claims in public that the company manipulated their machines to benefit Joe Biden caused damage to their reputation and business (Grynbaum & Bromwich, 2021).

Potential economic harms can be felt on an individual level with scam artists using misinformation like fake job adverts or impersonating bank communication in phishing attempts to steal money or personal details (Cunliffe-Jones, 2021, MediaNumeric interview).

6.7. Fact-Checking: History, Process & Skills

An unprecedented rise in the scale and reach of misinformation in the digital age has been met with a heightened media focus on the problem and a rise in fact-checking organisations dedicated to stemming the rising tide. There are now over 300 active fact-checking organisations spread across every continent in the world (Mark & Luther, 2020). Ninety-two of these organisations are accredited members of the International Fact-Checking Network (IFCN) and are signatories to the IFCN's code of principles to ensure cohesion and best practice (Orsek, 2021).

Fact-checking organisations are united in the goal of addressing the increasing flow of misinformation across online and offline networks by making available accurate information on important topics. Fact-check articles seek to verify true facts, debunk misinformation and place claims in context without which they may be misleading. Through a partnership programme with Facebook independent fact checkers attach labels on false or misleading content to warn users and reduce its distribution in the news feed. In addition to fact-checking some organisations promote media literacy in the community to empower people to spot misinformation for themselves and work with institutions to advocate for better access to public data and policy outcomes in an effort to prevent or remedy the broader problems of misinformation. All these activities are focussed on increasing the supply of good information and finding ways to limit the supply and impact of bad information.

6.8. Beginnings & Recent Trends

The practice of fact-checking political speech and election campaign literature started in the 2000s with projects like Factcheck.org that the Annenberg Public Policy Center at the University of Pennsylvania launched in 2003. Factcheck.org's mission, as Eugene Kiely, the organisation's director since 2012, explains, is to be a source of accurate information to the public and the media with the goal of reducing the level of deception in US politics (Kiely, 2021, MediaNumeric interview).

Since that beginning two moments have seemed particularly significant to the growth of this journalistic practice. A first wave of growth was inspired by the 2009 Pulitzer Prize awarded to PolitiFact, the fact-checking project launched in 2007 at what was then the *St. Petersburg Times* (now *Tampa Bay Times*). PolitiFact's innovation was their Truth-O-Meter, a six-point rating system that introduced nuance, clarity and structure to evaluations. A second wave was sparked by the global deluge of 'fake news', fabricated or hoax stories online using presentation to masquerade as serious news, which gained prominence in 2016 (Mantzarlis, 2018). The analysis by BuzzFeed's Craig Silverman showing that the top-performing fake news stories related to the 2016 US president election on Facebook generated more engagement than the top stories from major news outlets sounded the alarm about the power of social media to amplify misinformation and assist its dissemination across the world (Silverman, 2016).

While a global surge of misinformation has been facilitated by the reach of online platforms, so too has the digital expansion of the last decade given fact-checking organisations access to sources of information and tools necessary to counter it; the internet has made possible a global explosion in fact-checking (Kessler, 2020). According to the Duke Reporter's Lab first annual census in 2014, there were 44 active organisations clustered mostly in Europe and North America. The Duke study found that around half of them were affiliated with legacy media organisations like newspapers and television networks and the other half were non-profit groups engaged in a form of public interest journalism (Adair, 2014). By May 2021 active fact-checking organisations had exploded to 307 operating in over 80 countries representing every continent across the globe, according to

Duke's running count. While organisations continue to rise steadily in Europe and North America the recent numbers have been driven by considerable growth in Asia, particularly in India, with the number of fact checkers more than doubling since 2019 (Mark & Luther, 2020).

Prompted by the burgeoning numbers of fact checkers all over the world the International Fact-Checking Network (IFCN) was founded at the Poynter Institute in 2015. Soon after the IFCN introduced its Code of Principles to promote minimum standards in its verified signatory organisations. As IFCN director Baybars Orsek explains, with the introduction of the Code of Principles the IFCN started to promote higher standards in fact-checking particularly around transparency, accountability, journalistic principles such as corrections policy and displaying a solid methodology on the organisation's website. More recently the IFCN has started providing resources through grant programs, exchanges, fellowships and tool bags; different resources for organisations to increase their skills and knowledge and invest in their sustainability (Orsek, 2021, MediaNumeric interview).

Recent trends have seen fact-checking morphing from a predominantly non-profit public interest journalism project to a for-profit exercise (Orsek, 2021, MediaNumeric interview). The Poynter Institute *State of Fact-Checking 2020* report notes a dramatic rise in for-profit fact-checking organisations from constituting around a third of all operations to over half between 2018 and 2020: "for-profit fact checkers are a driving force within the community" (p. 2). Baybars Orsek puts this down to the popularity of fact-checking and recent investments in the community. The vast majority of fact checkers report receiving most of their funding either from donations, memberships and grants, or from the most recent and now dominant source of funding, representing the main source for 43.06 percent of organisations, the Facebook Third-Party Fact-Checking Program (Poynter Institute, 2020).

6.9. The Fact-Check

Fact-checking organisations are focussed on increasing the supply of good information and finding ways to limit the supply and impact of bad information. The principle weapon in the armoury of organisations to challenge misinformation has been developed by what Amy Sippitt and Will Moy (2020) refer to as "first generation" fact checkers, the post hoc journalistic fact-check (p. 592).

Sippitt and Moy (2020) see three benefits to fact-checking: providing accurate information on important issues for people to make up their own minds; holding public figures or instructions to account for the inaccurate or misleading claims they might make; building an evidence base of unsubstantiated claims and how they arise.

For Baybars Orsek (2021, MediaNumeric interview) all fact checkers are journalists and fact-checking is primarily a journalistic practice, but a fact-check article differs from he-said-she-said journalism that seeks balance. PolitiFact's editor-in-chief Angie Drobnic Holan explains "the mission of fact checkers is to document what is real and what is not; what is verified

and what is not; what is provable and what is not" (Holan, 2021, MediaNumeric interview). When they call something a fact-check they make two promises to their audience: that they are looking at a matter of fact, not opinion, and that they are going to have a finding at the end.

To do this, fact checkers follow a method that can be broken down into three basic stages: find a factual claim to check; verify or debunk the claim with reference to facts and evidence; and correct the record by pushing a considered and well-evidenced verdict (Mantzarlis, 2018).

6.9.1. Finding the Claim

Most fact-checking organisations will use a transparent method to identify the claims that they are going to fact-check. Most of these methods consider issues like:

- Is it an important issue?
- Is it a significant claim?
- Is it a statement of fact that can be checked?
- What harm could be caused if the claim was left unchallenged?
- Is the content trending and being engaged with by lots of people?
- Is the statement likely to be passed on to others?
- Has a claim like this or from this source been fact-checked before?

Some take tips or accept questions from an interested audience of readers, like FactCheck.org's "Ask FactCheck" feature.

Many monitor sources manually, building lists of pages, groups and subjects of interest. Others use social media analytics tools like TweetDeck or CrowdTangle to determine what's trending online as a starting point to finding important claims to check.

When identifying trending claims on social media fact checkers see two types of virality: a single claim being shared or viewed a lot or multiple instances of the same or similar claim each gaining a small amount of engagement.

At the heart of all these activities and criteria is the recognition that with finite resources and overwhelming quantities of information flowing over multiple surfaces the choice of claim to check is important and can be summed up with two questions: can it be fact-checked and should it be fact-checked?

6.9.2. Monitoring Social Media: CrowdTangle

CrowdTangle is a tool that fact checkers use to monitor and analyse social media content. The tool tracks public content on the most popular Facebook accounts and groups, Instagram and reddit and reports who is posting content, what type of content it is, which accounts have shared it and crucially how many interactions, measured in likes, reactions, comments and shares, the content has gathered.

Fact checkers build lists of pages and groups they want to monitor and use CrowdTangle to sign post them to the most viral content: “CrowdTangle has been a huge help with some of our Facebook fact-checking because it just shows us what's most popular” (Holan, 2021, MediaNumeric interview).

Other tools that fact checkers use to monitor and analyse content in online spaces include: Trendolizer, a tool for monitoring the virality of images, videos and stories; TweetDeck, which provides a dashboard to manage Twitter accounts, monitor curated searches and analyse Twitter content; and BuzzSumo, a paid for monitoring service that identifies what is trending across different online networks.

6.9.3. Finding the Facts

According to Eugene Kiely “the number one rule of fact-checking is to check with the source” (2021, MediaNumeric interview). This is a basic journalistic tenet that most fact checkers agree upon: go to the source of the claim and ask them with what evidence they are backing up their claim. The onus is on the one making the claims to back it up.

Evidence is evaluated with reference to reliable repositories of public data, like government statistical services or the World bank for example. Researcher and Africa Check founder Peter Cunliffe-Jones explains that access to reliable data is critical to fact-checking efforts and lacking in some parts of the world, which is why Africa Check set up their searchable Info Finder database, which recommends the best sources of information on a range of topics for a range of African countries (Cunliffe-Jones, 2021, MediaNumeric interview).

Expert opinions can be sought to provide further context and perspective on the claim but Peter Burger of Leiden University and Nieuwscheckers cautions that it is important not to rely on expert opinion alone. Fact-checking is an evidence-based practice: “It shouldn't be a matter of one authority saying this is bunk. They have to prove it” (Burger, 2021, MediaNumeric interview). Fact checkers also use an array of online tools to verify or debunk content, especially claims involving potentially manipulated photographs or videos. Innovative tools like the InVid-WeVerify plug-in analyses the content of images and aggregates similarity searches across multiple search engines to detect whether it has been manipulated.

6.9.4. Verifying Content: InVid-WeVerify

The InVid-WeVerify plug-in started as a tool to verify UGC images and videos and then pivoted to help fact checkers and journalists to tackle misinformation. The freely available plug-in has proved popular attracting to-date around 44,000 users. It assembles a diverse suite of verification tools enabling fact checkers to analyse the content and context of pictures and videos in detail.

The tool facilitates reverse images searches accessing seven different search engines including Google, Yandex and Baidu. For videos it automates the process of splitting it down into keyframes that can be searched for by similarity to see the contexts they’ve been used in before. It also has a

facility to check the metadata of videos and pictures which may contain important clues as to where and when the image was produced.

For content analysis InVid-WeVerify offers a lens tool to enhance sections of an image, an optical character recognition (OCR) tool to extract text information from inside a picture, like the wording on a banner at a demonstration and various forensic filters to search for clues of inconsistencies and manipulations.

Lastly, InVid-WeVerify provides a suite of search tools, monitoring and facilities to analyse tweets to estimate their veracity based on user comments and an advanced tool to search across networks, for example, to search for an hashtag found on one social media platform to see whether it is trending on popular platforms like YouTube, Twitter, Facebook, VK, 4Chan or 8Chan and also on underground networks.

As head of AFP's Media Lab Denis Teyssou, one of the developers, explains the toolbox is evolving all the time, adding and improving functionalities and appearance with a strong emphasis on being user-friendly. It is aimed not only at fact checkers, but other journalists, media literacy researchers, human rights defenders and citizens (Teyssou, 2021, MediaNumeric interview).

6.9.5. Correcting the Record

The principal format that fact checkers use to present their results is an article published online. Fact-checking experts we interviewed highlighted a number of different elements that make a good fact-check:

6.9.6. Speed

Many identified speed as one of the most crucial factors; fact-checks need to be made in a timely manner. The longer a false claim is left to circulate unchallenged the more people will see it and the more damage it can do.

6.9.7. Accuracy

Fact checkers need to work quickly but it is imperative that their work is accurate to maintain credibility. For this reason Eugene Kiely says that all FactCheck.org's stories are fact-checked internally before publication (2021, MediaNumeric interview).

6.9.8. Tone

Another element that is important is tone. Fact checkers present information to a sceptical public in a way that allows them to accept the facts as they are presented: "You don't want to use loaded language; you don't want to use unnecessary adjectives; just the facts, ma'am, as a TV show used to say" (Kiely, 2021, MediaNumeric interview).

6.9.10. Context

Not all claims that fact checkers scrutinise are out-and-out falsehoods. Many of them have a kernel of truth that makes them convincing but that kernel of truth is taken out of context which makes the claim misleading. Peter Cunliffe-Jones explains that:

> "Very often the reason that people cling to a belief when it's been shown to be untrue, is that there is an element of the claim that is true, and so they cling to whole misbelief." (Cunliffe-Jones, 2021, MediaNumeric interview)

So, the best way to correct somebody's false belief is to help them come to that themself by acknowledging the element of the claim that is true but then place it in its proper context. In this way "you deal with the factor that is making them cling to a false belief by putting it in context" (Cunliffe-Jones, 2021, MediaNumeric interview). Including details about when and where similar claims have surfaced before is also another way in which the attraction of a false claim can be weakened by contextualising it (Cunliffe-Jones, 2021, MediaNumeric interview).

6.9.11. Transparency of Sources

Giving people access to good information to help them come to the facts themselves means giving people access to all the information. It is important that fact checkers are not only transparent about the sources they use, but also that they provide public links to that data so that any reader can look up the data for themselves and fact-check the fact checkers (Kiely, 2021, MediaNumeric interview).

6.9.12. Clarity of Conclusion

Fact checkers attempt to take into account the full complexity of a question and deal with all its nuance, but in the end fact-checks work best when they arrive at a clear conclusion. The claims they deal with are rarely simply 100 percent true or false; fact checkers need to deal in the shades of grey in between. For this reason some organisations have developed their own rating scales. The pioneer in this is PolitiFact with their Truth-O-Meter that rates claims as: True; Mostly True; Half-True; Mostly False; False; and Pants on Fire. A lot of organisations use similar ratings scales with associated graphics of Pinocchio (*The Washington Post* fact checker) or a Bloodhound (Animal Político's El Sabueso). Such scales allow fact checkers to easily communicate nuance in their conclusions in an attractive way that is easy to share online (Funke, 2019). Although some fact checkers do caution that such ratings introduce an element of subjective interpretation in a manner that could be interpreted as confrontational which might alienate those readers fact checkers most need to reach (Krueger, 2017).

6.9.13. Distribution

One hundred percent of fact checkers surveyed by the Poynter Institute for their *State of Fact-Checking 2020* report, reported that they distribute their work online, although 24 percent reported that they still use TV and 19 percent print. While the online fact-check article remains the

principal means by which fact-checking organisations publish their work, organisations are diversifying their engagement strategy to vary their offer.

A project run by the Italian fact checker Pagella Politica, The Fact-Checking Engagement Project, looked at alternative innovative means through which fact checkers around the world engage their audience (Loguercio & Canepa, 2021).

Their report details that fact checkers are:

- Using techniques of digital marketing to build brand identity through images and graphics, like PolitiFact's Truth-O-Meter, on social media. Images or graphics can present simple verdicts on claims or signposts to full articles.
- Distributing digests and recaps in the form of a regular newsletter, which can signpost to the organisations social media channels and full articles.
- Engaging their audience with quizzes and games that draw upon information in regular fact-checks. An example of this is ColombiaCheck's "Coronaquiz" which presents readers with questions relating to dubious claims about the pandemic. Each question links back to a more in depth fact-check article.
- Speaking directly to their audience using podcast formats ranging from the Full Fact Podcast which presents a 30-minute deep dive into a single subject, weekly round-ups of the most relevant debunks like Maldita's podcast or a single short audio report detailing presenting a verdict on a single claim like that of Portuguese news outlet Observador.
- Using short-form videos on social media platforms to bring their message to their audience. Some, like PolitiFact, have a dedicated YouTube channel.

6.9.14. ClaimReview

Once a fact-check article is written it is generally published on the website of the organisation that produced it. This serves a ready made and interested audience that visits the organisation's website looking for this information.

To make fact-checks travel further and reach other audiences a schema, a kind of labelling system, called ClaimReview was designed as a standardised way to describe fact-checks and to help search engines find them. ClaimReview is the structured mark-up that sits behind a fact-check and flags them to platforms like Google, Bing and Facebook and their products: "Describing fact-checks consistently, as a specific type of content with inherent structure that is universally understood by fact checkers and distribution platforms, is vital for fact-checking to operate at internet scale" (Full Fact, 2020c, p. 57).

Google, who is part of the collaboration that maintains ClaimReview, announced in 2019 that the schema had helped fact-checks appear more than 11 million times a day in Google Search results globally and in Google News in five countries (Brazil, France, India, UK and US); this makes roughly 4 billion impressions a year (Mantzarlis, 2019).

According to Andy Dudfield, head of automated fact-checking at Full Fact, the library of fact-checks curated using the ClaimReview schema has reached around 100,000 (2021, MediaNumeric interview).

6.9.15. Media Partnerships

Sippitt and Moy (2020) point to media partnerships as another way in which fact checkers can significantly increase their reach. As part of their strategy for reporting on the 2019 UK general election Full Fact reached millions of people with their fact-checks through partnerships with Sky, the *Evening Standard* newspaper and TalkRadio. Similarly, fact checkers regularly syndicate their content to other media organisations to increase exposure. FactCheck.org's fact-checks can be found on MSN and elsewhere for example (Kiely, 2021, MediaNumeric interview).

6.9.16. Social Media Fact-Checking

While social media platforms have recently emerged as a main conduit of misinformation some platforms are working with independent fact checkers to tackle the flow of false and misleading content. Platforms like Twitter, Pinterest, YouTube do have their own policies and community standards on misinformation. WhatsApp lists the contact details of 48 fact-checking organisations in 28 countries on their website for users to check suspect content with. Facebook, however, engages directly with independent fact checkers and "is the only internet company with a robust global programme to tackle misinformation and a mechanism for labelling and acting against false claims" (Full Fact, 2020c, p.81).

The Facebook Third-Party Fact-Checking programme now includes 80 independent fact checkers who review content in more than 60 languages according to a recent article by Facebook (Rosen, 2021).

6.9.17. How Does it Work?

Facebook uses suggestions from users and machine learning tools to surface and collate content that may be misleading into a queue. fact checkers can access this feed through an interface provided by Facebook and rate the claims and attach their fact-checks. Claims can be rated as: False, Altered, Partly False, Missing Context, Satire and True. It is important to note that Facebook surfaces the content for checking but it is the independent fact checkers who give the rating without input from Facebook.

Once the content has been rated Facebook can then take one or more of the following actions: apply notices that pop up when users try to share or have shared fase-rated content; apply misinformation warning labels; reduce the content's distribution in their News Feed and other surfaces; sanction users or websites that repeatedly share false-rated content. These actions appear designed to discourage the sharing of misleading content, to reduce the impact of false-rated content so that fewer people see it and apply warning labels and further information

provided by the fact checkers for those that do see it. Facebook only removes content entirely if it breaches their community standards and this is a determination that the company makes.

One notable exception is that Facebook does not allow fact checkers to rate speech from politicians, neither their organic content nor political ads. In explanation Facebook says that they do not think it is appropriate for them "to referee political debates and prevent a politician's speech from reaching its audience and being subject to public debate and scrutiny" (Clegg, 2019). The recent Facebook Oversight Board decision to uphold the platform's indefinite suspension of former US president Donald Trump over supportive messages of his supporters' siege of the US Capitol has however renewed calls from the fact-checking community for a review of the company's stance (Mantas, 2021).

With Facebook's use of machine learning to collate the feed of potential misinformation and markers that help face checkers to identify pieces of content that are most viral, Angie Drobnic Holan feels that the programme "has just been a game changer as far as our ability to identify misinformation online" (2021, MediaNumeric interview). The input from Facebook makes misinformation trending on their platform easier to find and reduce its impact. And the programme also facilitates exposure for fact checkers' work to the global audience of a social platform with over two billion worldwide users (Full Fact, 2020c).

The Facebook Third-Party Fact-Checking Program has also funded the growth and development of global fact-checking. As Sophie Nicholson from AFP's Fact Check team says: "A big part of why it's grown so fast, at AFP and elsewhere, is because Facebook has been financing fact checkers" (Nicholson, 2021, MediaNumeric interview). As already mentioned 43.06 percent of all fact checkers surveyed by the Poynter Institute identify Facebook as their principal source of funding (2020).

Peter Cunliffe-Jones feels that this level of cooperation and monetisation with social media platforms has led to a general shift within the fact-checking community toward placing a much greater focus on online misinformation (Cunliffe-Jones, 2021, MediaNumeric interview). Researchers Lucas Graves and Alexios Mantzarlis also recorded a shift in focus toward viral rumours despite organisations having the mission to target political lying (2020).

Taking account of the benefits and with a need for more frameworks on policing online misinformation in mind Angie Drobnic Holan would like to see other social media platforms emulate Facebook's Third-Party Fact-Checking programme (2021, MediaNumeric interview). A view shared by Full Fact in their 2020 report on their experience with the programme (Full Fact, 2020a)

6.10. Automated Fact-Checking

Fact-checking is time-consuming and hard. In today's information environment fact checkers need to filter overwhelming amounts of information to find huge numbers of claims and then decide which to work on; they need to engage with hard-to-reach target audiences across different platforms; they need to operate at internet scale and they need to work fast. To help with these challenges, as Andy Dudfield explains, fact checkers are working with tech companies and each other to develop artificial intelligence tools (Dudfield, 2021, MediaNumeric interview).

The current goals are to try to find parts of the fact-checking workflow that AI can automate to save fact checkers time and energy. The field of automated fact-checking however has entered the sector's lexicon as a misnomer that has caused confusion. Some fact checkers are sceptical about the possibility and desirability of replacing human intelligence and intuition with machines. However, the goal of automated fact-checking is not to replace human fact checkers, the goal is to provide automated tools to assist and augment their work.

There are currently three broad areas of the fact-checking workflow where automated tools are being developed and deployed:

6.10.1. Scale: Monitoring & Identifying Claims

The first part of the fact-checking workflow that is being assisted by automated fact-check tools is at the start of the process helping fact checkers identify the right claims to check. Rather than a human fact checker listening to hours of parliamentary debate, for example, an AI model can be trained to pass through all that information and to reduce it down to just those things that seem to be claims. These claim-like statements can then be structured, classifying them with additional information; who has said them, what type of claim it is, where it came from, etc. This produces a more nuanced and important collection of information for the fact checker to work with and make the ultimate decisions about.

6.10.2. Speed: Speeding up the Checking of Statistical Claims

The second area where automated fact-checking can augment the efforts of fact checkers is in speeding up the process of checking a claim. This can mean having good transcription tools to turn speech into searchable text that can then be analysed or having dashboard tools displaying high quality data all neatly laid out. Where high quality publicly available machine-readable data exists it should be possible to start to automate some parts of the fact-checking process. For instance, where an AI model identifies a statistical claim, say GDP is rising, it can automatically access the corresponding National Statistical Institute's API and test that statement against the official data. A tool being developed by Full Fact aims to do precisely this and is trained on 15 topics with around 60 verbs defining trends (e.g. rising, falling). The tool can then lay all the data out neatly and report it back to the fact checker. This is particularly useful for live fact-checking at events, like a live debate, or a radio show, for example, where fact checkers need real-time information. The challenges of this emerging field lie firstly in the complexity of language, there are many different

ways in which these kinds of statistical claims can be phrased. A second and more pertinent challenge is the paucity of high quality openly published statistical information that can be relied upon to support this process.

6.10.3. Impact: Claim Matching

A third process where AI is particularly helpful is in finding repetitions of a claim. Claims can be repeated multiple times and surface across different platforms. So, once a claim is fact-checked AI can help identify when and where that claim is being repeated so that the fact-check can be applied to all multiple instances of the claim. This makes the most of the impact of each fact-check.

6.11. Required Skills & Knowledge

6.11.1. Journalistic Skills

All fact checkers are journalists. To begin with fact-checking requires those basic journalistic skills, people skills, communication skills, investigative skills and editorial judgement. All fact checkers are journalists first and foremost and need the skills of a journalist (Orsek, 2021, MediaNumeric interview).

The first simple journalistic skill and the number one rule of fact-checking is to always seek out the source. Whether it is a political claim in a speech or a viral photograph surfacing on social media, finding the source is the most important thing a fact checker can do. For this a fact checker needs good search skills to know where to look to find the relevant information. (Kiely, 2021; Cunliffe-Jones, 2021; Teyssou 2021; Holan, 2021, MediaNumeric interviews)

An important journalistic value for fact-checking is to approach every claim with an open mind, not rush to judgement and remain impartial. Fact checkers are often accused by elements who find their articles unwelcome of being partisan. Being aware of the types of thinking biases that all people have can be important for understanding the fact checker's own reaction to the claim in front of them and help them to remain impartial and focus on gathering all the facts (Cunliffe-Jones, 2021; Kiely, 2021, MediaNumeric interviews).

Fact checkers need investigative skills. They need to do research and access primary sources. They need good search skills and an understanding of the online environment to know where to find the relevant information and strong reading comprehension skills to take it in and synthesise (Kiely, 2021; Holan, 2021, MediaNumeric interviews).

Fact-checking involves speaking to people, checking details and gathering the opinions of experts. Communication skills are very important. And at the other end of the process they need writing skills to be able to present difficult ideas in a clear and engaging way; Angie Drobnic Holan said that the one place in which a lot of applicant fact checkers fall down is in their ability to synthesise all the information they have found and write the story (2021, MediaNumeric interview).

6.11.2. Knowledge Drawn from Digital Literacy Training

Developed through studies into digital media literacy training programmes, fact checker and researcher Peter Cunliffe-Jones and colleagues have developed the six Cs — six knowledge areas that can help fact checkers understand and identify misinformation (Cunliffe-Jones et al., 2021b). These are:

- **Context** - knowledge of the context in which you're likely to see misinformation. So, if there is a disaster, if there's a bomb blast in your town, be aware that you are very likely to see much misinformation about that or a flood or a forest fire or whatever it might be.

- **Creation** - understanding who the sources of misinformation are and how it is created. Understanding how to distinguish between systems of information creation that try to promote accuracy and those that don't: credible news organisations versus hyper partisan sources or junk news websites, for example.

- **Content** - how to analyse the information conveyed in a claim or the content of an image. This is about understanding the difference between facts and opinions or predictions about the future. Analysing if the content of a claim adds up or understanding that if a photograph of a natural disaster has been shown before then it can't actually have been from a recent event.

- **Circulation** - recognising the processes by which misinformation travels online and offline.

- **Consume** - understanding how people consume misinformation and why they or we ourselves might believe things that are wrong.]

- **Consequences** - recognising that misinformation has consequences. Reminding people of the potential consequences of misinformation has been shown to have a significant effect on whether people share misinformation. An appreciation of the potential consequences is useful to the fact checker to determine the importance of fact-checking a claim.

6.11.3. Specialised Skills

All fact checkers are journalists, but not all journalists are fact checkers. Journalists who are committed to fact-checking will need to invest time in learning more specialised and targeted skills in addition to those essential journalistic skills and values.

6.11.4. Lateral Reading

The way that fact checkers evaluate the credibility of unknown websites varies from the techniques used by lay people. A team of researchers from Stanford and Texas in the US wanted to see if they could improve college students' ability to critically evaluate online sources. They found that when faced with a known site most college students and professors would read the page

vertically from top to bottom, searching internally for clues in the text, the internal links, the About page and the URL. Details like a .org top-level domain, the layout and graphics and whether there are links to other websites will be taken, often erroneously, as markers of credibility. Using this method produced poor results when trying to assess the veracity and credibility of websites (Breakstone et al., 2021).

Fact checkers employ different techniques. When faced with an unknown website, they will leave the unknown website straight away, opening up tabs horizontally in their browser to consult other trusted sources on the wider web. Rather than reading the page vertically, the fact checkers used a technique called lateral reading. Using this method allows the fact checkers to identify the source of the known website, understand its associations with other organisations and more quickly and accurately establish the site's credibility. Teaching the college students lateral reading techniques significantly improved their performance on these types of critical analysis tests, providing empirical evidence that lateral reading is an efficient and powerful technique for a fact checker (Breakstone et al., 2021; Burger, 2021, MediaNumeric interview).

6.11.5. Technical Skills for Verifying Pictures & Video

To verify pictures and videos fact checkers need to be skilled in the techniques of endogenous and exogenous verification; that is the use of various tools to examine the content itself and tools to look at the context around the content (Teyssou, 2021, MediaNumeric interview).

Endogenous verification is analysing the content of the image. It could be visual verification, enhancing the photograph to look a little deeper at the content or viewing a video frame-by-frame looking for inconsistencies or to isolate details, like street signs, to verify the location. Fact checkers can also use forensic tools and filters to look at the content differently to find indications that it has been altered. Other techniques could be to verify the time of day by using tools to measure the angle of the shadows (Teyssou, 2021, MediaNumeric interview).

Exogenous verification techniques look beyond the content at the context around the image or video for clues as to its veracity. Here you may be looking at who is talking about the content and what they are saying, who shared it and for what purpose. To find the source of an image or video shared online fact checkers can use tools to search by similarity — reverse image search — to see if the same or similar image has been used before in a different context. Comparison with earlier usages of the image can reveal if something has been added or taken away or, as is often the case, if it is an undoctored image of a previous event being used out of context. Another common technique to verify the location depicted in an image is geolocation using Google Maps (Teyssou, 2021, MediaNumeric interview).

6.11.6. Content Knowledge

Fact checkers' attention moves over different subjects that carry their own different requirements of knowledge to understand. Like any journalist wanting to report on a particular specialism, fact checkers require a good grounding in the field that they wish to investigate.

Political fact checkers need to have that grounding in politics and government; how the government operates, how the hierarchy of elected figures and appointed officials works, and knowledge of the sources of official parliamentary business.

The Covid-19 pandemic has necessitated bringing medical reporting into the mainstream for fact checkers. Dealing with misleading health advice, medical hoaxes and scientific misunderstandings requires a basic grounding in the science involved and the ability to communicate often technical ideas in a clear and concise way. A familiarity with the technical literature is also an advantage; Peter Burger teaches his students how to read the standard format of scientific papers and medical reports (Kiely, 2021; Holan, 2021; Burger, 2021 — all MediaNumeric interviews).

6.11.7. Data Skills for Verifying Statistics

Statistical claims are very common forms of political speech. Rarely is a statistical claim totally false, more commonly there is some kernel of truth to it; it often comes down to how you look at the data. That's why a basic knowledge of statistics is an essential weapon in the armoury of the fact checker, says machine learning expert at the Polish- Japanese Academy of Information Technology Bartlomiej Balcerzak (Balcerzak, 2021, MediaNumeric interview).

Fact checkers need to know where to find relevant and reliable data, and to know what to do with it when they've found it. They need to know what conventional measures are used. If, for example, they are trying to verify a claim on the issue of poverty, they need to know what the best measure is in the official data; is poverty measured by average household income or median earnings? Is it an absolute measure or a relative measure? What measure is the claim using? Is it the best measure? In this way knowing the strengths and limitations of the data before you will also give the fact checker an indication of the ways in which that data can be manipulated or misinterpreted (Balcerzak, 2021, MediaNumeric interview).

6.12. Fact-Checking: Evaluation, Successes & Challenges

6.12.1. Does Fact-Checking Work?

How successful is fact-checking at reducing misinformation and the misperceptions that it causes?

Since approaches are varied and multifaceted there are many metrics by which success could be measured. In what follows we will consider some of the themes that stand out. This is not meant to be an exhaustive systematic review of all available evidence, as such an undertaking is beyond the scope of this report.

It is recognised that it is hard to change people's strongly held views. Everyone has thinking biases and the facts of the matter are not always the only factor involved in the formation of beliefs (Kahneman, 2011). People can resist factual information that goes against their pre-existing

attitudes or favour the conclusions that fit most comfortably with their world-view, which can lead people to disagree on what the facts really are (Kunda, 1990).

Early research seemed to support the idea that counter-attitudinal fact-checks struggle to change people's minds, even sometimes reinforcing misperceptions in what was dubbed the "backfire effect" (Nyhan, 2010).

A recent review of studies however found that observation of a backfire effect was rare and the cases where the effect was found tended to be associated with particularly contentious topics, or where the factual claim was ambiguous (Sippitt, 2019). In a series of five experiments Ethan Porter (2017) also found no evidence of backfire following corrections on polarising issues finding that citizens generally heed factual information. Brendan Nyhan, the author who originally identified the backfire effect, has most recently suggested that the finding "appears to have been anomalous" (Nyhan 2020, p.231). Sippitt concluded that fact-checks help to inform the views of citizens (2019).

A different meta-analysis on fact-checking and corrective information similarly finds that fact-checks generally increase the accuracy of people's beliefs and reduce misperceptions (Walter & Tukachinsky, 2019).

For example Porter et al. (2021) found that exposure to fact-checks targeting Covid-19 vaccine misinformation were effective at reducing false beliefs about vaccines and counteracted the negative effect of misinformation on belief accuracy. Participants who encountered a correction after misinformation were found to be approximately as accurate as subjects in the control group who did not encounter misinformation. Exposure to fact-checks increased the accuracy of participants' beliefs but this did not translate into an effect on vaccination intention. The conclusion here is that updating beliefs about vaccines did not change behaviour intention accordingly.

This supported a similar finding in experiments on fact-checking claims made by Donald Trump during the 2016 US presidential election. Exposure to fact-checks prompted participating Trump supporters to update their factual beliefs but had no effect on attitude toward Trump and did not affect their voter preference (Nyhan et al., 2019).

So, evidence suggests that exposure to corrective factual information can increase the accuracy of people's beliefs on important subjects, but this does not necessarily or immediately translate into a change in behaviour.

But what about the public figures and institutions whose claims are held to scrutiny; is there any evidence that fact-checking deters public figures from making false or misleading claims?

6.12.2. A Culture of Accuracy

Although important and at the heart of what all fact checkers do, Full Fact believes that it is not enough to just make good information available and hope to nullify the harm of misinformation and correct the public's inaccurate beliefs (Sippitt & Moy, 2020). Following up with the person or organisation that made a misleading claim to request a correction is seen as a key part of the process to encourage behaviour change, promote a culture of accuracy within institutions, organisations and public officials and disrupt the supply of false and misleading claims (Full Fact, 2021).

While in office, former US president Donald Trump was one of the most prodigious and high-profile sources of misinformation. He was also one of the most relentlessly fact-checked sources of information through his four years in office. Despite this unprecedented personal attention *The Washington Post*'s Fact Checker team that catalogued Trump's false or misleading claim through his White House tenure closed the count at over 30,000 claims. From this evidence alone you could be forgiven for concluding that aggressive fact-checking does not necessarily change the behaviour of people in power.

However, Glenn Kessler (2020), head of *The Washington Post* Fact Checker team, notes in his book *Donald Trump and His Assault on Truth* that during the 2020 presidential campaign Trump's Democratic rivals took actions like dropping talking points and issuing apologies to avoid receiving negative fact-checks. Sippitt and Moy (2020) relate similar anecdotal evidence from a UK government official who told them that one of their key performance indicators was to avoid criticism by Full Fact. So, in some cases it seems possible that fact checkers could foster a culture of accuracy as public officials seek to avoid criticism.

Empirical research provides support for the idea that fact-checking can act as a deterrent. One field experiment found that politicians, in this case state legislators in the US, who received letters explaining the risk to their reputation if they were caught making false or misleading claims were substantially less likely to go on and receive negative fact-checking ratings or have their accuracy questioned. The authors suggest that the threat of negative ratings provided an incentive to the legislators that improved the overall accuracy of their public statements (Nyhan & Reifler, 2015).

Over the course of 2020 Full Fact sought 161 corrections, with 72 of these being successfully resolved; this is compared with 126 interventions, of which 51 were fully resolved, in 2019 (Full Fact, 2021). Although Full Fact did express some frustration at the length of time that some corrections took to be resolved, leaving the misleading claim in circulation where it can be picked up, reproduced or amplified. Furthermore, they point to the need for better ways for politicians to correct the official record at the UK parliament.

Fact-checks can sometimes prompt public institutions to radically change their stance. Researcher and Africa Check founder Peter Cunliffe-Jones relates that in 2018 the South African police published their crime figures for the year. Africa Check looked at the figures and worked out that they'd miscalculated every single crime statistic in their annual report because they'd taken a false

figure for the overall population. The miscalculation had the effect of diminishing rates of crime on the upward trend, making them not appear so bad, and deepening falling rates of crime making the decrease appear more dramatic. Following an analysis from Africa Check the police revised their crime rates to correct the error (Cunliffe-Jones, 2021; Wilkinson, 2018).

Having policymakers officially correct mistakes is critically important since distortions in the data can lead to accordingly skewed policy. Allowing policymakers to exaggerate their performance or diminish the performance of their opponents undermines proper accountability. Bad information can lead to bad policy (Cunliffe-Jones, 2021).

This is why what Full Fact refers to as “second generation” fact-checking organisations seek to engage with insinuations to secure systemic changes to limit the supply and impact of information (Sippitt & Moy, 2020).

6.13. Outreach & Advocacy Activities

While all fact-checking organisations find claims to check and publish their articles online and elsewhere, some organisations approach the issue of bad information as what Peter Cunliffe-Jones called a “whole society problem” (Cunliffe-Jones, 2021, MediaNumeric interview). These organisations are what Sippitt and Moy (2020) refer to as the ”second generation” fact checkers, the non-governmental organisations like Full Fact, Africa Check and Maldita in Spain for example.

These organisations place as much emphasis on outreach, advocacy and community education as they do on publishing fact-checks. They approach the bad information problem from the supply-side and work with institutions, policymakers and in the community to find solutions.

For IFCN director Baybars Orsek Full Fact is one of the leading organisations in the fact-checking community in terms of their commitment to advocacy (Orsek, 2021, MediaNumeric interview). Full Fact’s activities directed toward the UK government are wide-ranging and they seek policy solutions to address the broader problems of bad information and encourage a culture of accuracy in institutions and politics; they provided MPs with regular briefings on their work tackling misinformation relating to the novel coronavirus pandemic; and their 2021 annual policy report offers recommendations and calls to action to improve access to and quality of government information and to make sure that data is used correctly. Currently Full Fact are running a consultation toward developing a framework for addressing information incidents. This project seeks to develop a common assessment system, perhaps resembling something akin to military alarm levels, and a framework for action to overcome predictable difficulties in the distribution of reliable information or the tackling of misinformation on a global level; incidents like a global public health crisis (Full Fact, n.d.).

Both Africa Check and Maldita are examples of organisations with strong training arms. Both organisations run training courses in verification and media literacy to educate people in their

communities and empower them to spot misinformation for themselves. Africa Check has worked with over 5,000 individuals from the media, civil society, government, the private sector and the general public. In addition to training, Africa Check conducts research and works with partners across the continent to develop the Info Finder tool which is a resource that directs users to the best sources of information for a given subject (Maldita.es, 2018; Africa Check, n.d.).

6.14. Challenges

6.14.1. Funding

The first challenge in the minds of many fact-checking organisations is funding; how to keep their organisation going. Most of the 92 verified signatories to the IFCN Code of Principles are small organisations, they are small because they are difficult to fund. According to Peter Cunliffe-Jones, a lot of people think the work of fact checkers is important, but not so many want to pay for it. (2021) Earlier in this report we noted that Facebook has emerged as the principal source of funding for the fact-checking community. There is general agreement from the fact checkers we've spoken to that heavy dependence on social media platforms or any one single source of insecure funding could have implications for the sustainability of the community. Put simply "it's not that healthy to rely on one provider or funder to keep up your operations." (Orsek, 2021, MediaNumeric interview)

6.14.2. Independence

The way you are funded and the transparency around funding is also important for maintaining independence and being seen to be impartial in a polarised political environment. Fact checkers can only be effective when they are seen as an authoritative and impartial source of information. Politically polarised environments all over the world make it even more important for fact checkers to be able to demonstrate their independence to their audience. One perennial problem is that it is rare in a country that all the political factions conveniently produce exactly the same amount of misinformation as the others. One or other faction usually has a greater propensity and more effective means to generate misinformation. If a fact checker chooses claims simply on the basis of the misinformation that is out there then they will inevitably be open to being accused of being partisan. (Peter Cuniffe-Jones, 2021, MediaNumeric interview)

6.14.3. Transparency

Furthermore, transparency and accountability does not mean the same thing in different countries and the diminished ability of fact checkers to be transparent about their activities in countries that are more hostile to journalists is a worry for Baybars Orsek (Orsek, 2021, MediaNumeric interview). The IFCN network has grown to 92 different organisations from 48 different countries, and there are organisations applying from countries as diverse as Pakistan, Turkey, Algeria, Azerbaijan, Bolivia and Myanmar. It's not that easy to apply the same principles around transparency and accountability and allowing for this diversity whilst maintaining the standards

and credibility of the network is a big challenge: "We are constantly rethinking, revisiting our criteria with our assessors with our advisory board members," Orsek says, "to make sure that fact-checking will not only be a practice for some newsrooms in some established and safe functioning democracies but can also function in countries with less freedom of speech and media" (Orsek, 2021, MediaNumeric interview).

6.14.3. Abuse

Even with best efforts to present impartial authoritative information in a transparent and non-partisan way, fact checkers are having to become used to harassment and online abuse. Respondents agree that in the European and US context at least this abuse appears to mostly emanate from the political far-right. Angie Drobnic Holan, editor-in-chief at PolitiFact, explains that "in the US especially, but it's probably in other places too, the far-right demonises fact checkers to try to undermine their legitimacy." Holan says that the level of hostility in the US to the media has radically increased during her time as a journalist (2021, MediaNumeric interview).

6.14.4. Uncertainty

A key challenge to those reporting on the global coronavirus pandemic and fact-checking the shifting landscape of guidance and advice is the need to deal with and communicate uncertainty. On a fast moving story like the pandemic and reporting on a disease about which more is being learned all the time where advice is changing it is challenging for fact checkers to establish the true facts. Another challenge is to communicate clearly the uncertainty around the facts; something that might be true today may not be in three weeks time. Communicating complicated science in a salient way that takes into account the necessary uncertainty but doesn't add to confusion is a challenge.

6.14.5. Access to Good Quality Information

A huge concern to the community, especially in the Global South, is access to and the quality of information. At a meeting of African fact checkers in 2017 this issue "was identified as the biggest single challenge by every single organisation in the room" (Cuniffe-Jones, 2021, MediaNumeric interview). While access to public data is comparatively much better in the UK Full Fact has recently published a report delving deeply into perceived deficiencies in how the UK government collects, uses and communicates information. Their analysis, using the coronavirus pandemic as a case study found, among other things, dangerous gaps in information about important sectors like social care, gaps in existing data, like about criminal courts, and data that should have been collected and wasn't, like on the supply and demand of personal protective equipment. On the government's use of data, Full Fact highlighted that government ministers repeatedly quoted statistics without making the source public, therefore making it very difficult for fact checkers to scrutinise their claims. Their 10 recommendations call for improvements in the collection, use and communication of information in the UK. (Full Fact, 2021). As Peter Cunliffe-Jones puts it, "I think that that question of the difficult journey of good information…is probably the biggest problem.

The problem that has most concerned people is misinformation" (Cunliffe-Jones, 2021, MediaNumeric interview).

Where good information exists it still needs to be in a form and format that can be used. People working on the next generation of automated tools and artificial intelligence inputs stress the need for high-quality well-published data that can be trusted. Andy Dudfield, head of automated fact-checking at Full Fact explains that for technology to really make a difference in the fact-checking process what is needed is good trustworthy data that is openly published with good metadata that is machine readable. It's not a government office publishing a PDF, a format that is time-consuming to work with, what is needed is data that is of the web and can be interacted with. There is some, but not enough of this data around and this is an area where improvements are needed (Dudfield, 2021, MediaNumeric interview).

6.14.6. Scale of Misinformation

Another challenge to fact checkers is the sheer scale and quantity of information flowing down the 'firehose of the internet'. As FactCheck.org's director Eugene Kiely puts it "the biggest challenge is just trying to get our hands around the vine of misinformation that exists" (Kiely 2021, MediaNumeric interview). In part this is a technical challenge, introducing workflows and tools that effectively filter vast quantities of information to find the claims fact checkers need to consider. How fact checkers level up to work at internet scale where the amount of false and misleading content easily outstrips fact checkers current capacity to publish fact-checks is a deep challenge for the future (Orsek, 2021, MediaNumeric interview).

6.14.7. Reaching the Right Audience

And once the fact-check is done, the final challenge is getting that information to the right audience. Although audiences for fact-checking output have grown considerably over the last 10 years, "if the audience who see the misinformation are not the same people as the audience who see the fact-check you're not actually doing much good in correcting false beliefs" (Cunliffe-Jones, 2021, MediaNumeric interview). For good information to have a positive effect it needs to be presented in the format that is most suited to the setting in which the misinformation was shared. Eugene Kiely sees the same problem, in the polarised political landscape of the US sections of the population seem to refuse to accept facts communicated from certain sources because they perceive those sources to be biased:

> "The only way to convince somebody who is kind of immune to our reporting is to have someone in their circle read it and share the information with them. So, my hope is just that by providing the accurate information for people who want to seek it out, then those people will then carry on that message to their circle of friends and family who might be sceptical receiving it from us, but would be more accepting of receiving that information from a trusted source… the biggest challenge I think is that, trying to reach that audience." (Kiely, 2021, MediaNumeric interview)

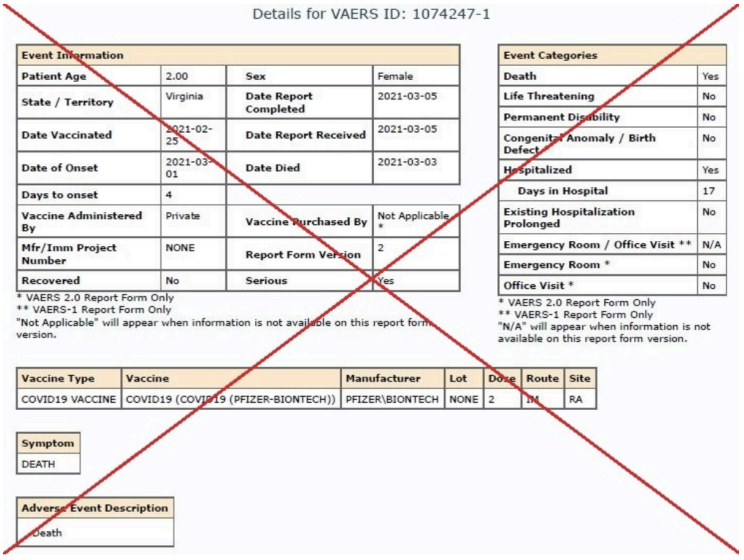
7. Where Data & Misinformation Collide

As developed earlier in this report, the Covid-19 pandemic has been a catalyst for both data journalism and data visualisations and the rapid spread of misinformation and disinformation.

Some of the misinformation, however, comes from open-source databases that are either being misunderstood or wilfully manipulated to serve an agenda. One particular database cited in much of the misinformation around Covid-19 is the Vaccine Adverse Event Reporting System[77] (VAERS) which is co-sponsored by the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA), both official agencies of the US Department of Health and Human Services.

Established in 1990, the VAERS is a national early warning system to detect possible safety problems in US-licensed vaccines. It is a passive reporting system, which means that it relies on individuals to send in reports of their experiences with vaccines to the CDC and the FDA. VAERS is not designed to determine if a vaccine caused a health problem, but it is useful for detecting a pattern of adverse side effects that might indicate a problem with a vaccine. Healthcare professionals are required to report certain adverse events and vaccine manufacturers are required to report all adverse events that come to their attention. However, anyone can report an event to VAERS and this makes the database open to abuse and manipulation.

In March 2021, a report was filed to VAERS that said that a two-year-old girl died in the state of Virginia less than a week after receiving a Pfizer-BioNTech shot.



Details for VAERS ID: 1074247-1

Event Information			
Patient Age	2.00	Sex	Female
State / Territory	Virginia	Date Report Completed	2021-03-05
Date Vaccinated	2021-02-25	Date Report Received	2021-03-05
Date of Onset	2021-03-01	Date Died	2021-03-03
Days to onset	4		
Vaccine Administered By	Private	Vaccine Purchased By	Not Applicable *
Mfr/Imm Project Number	NONE	Report Form Version	2
Recovered	No	Serious	Yes

* VAERS 2.0 Report Form Only
** VAERS-1 Report Form Only
"Not Applicable" will appear when information is not available on this report form version.

Event Categories	
Death	Yes
Life Threatening	No
Permanent Disability	No
Congenital Anomaly / Birth Defect	No
Hospitalized	Yes
Days in Hospital	17
Existing Hospitalization Prolonged	No
Emergency Room / Office Visit **	N/A
Emergency Room *	No
Office Visit *	No

* VAERS 2.0 Report Form Only
** VAERS-1 Report Form Only
"N/A" will appear when information is not available on this report form version.

Vaccine Type	Vaccine	Manufacturer	Lot	Dose	Route	Site
COVID19 VACCINE	COVID19 (COVID19 (PFIZER-BIONTECH))	PFIZER\BIONTECH	NONE	2	IM	RA

Symptom
DEATH

Adverse Event Description
Death

Figure 20: Screenshot of a since-removed report in the Vaccine Adverse Event Reporting System, as recorded by AFP Factcheck[78]

[77] Vaccine Adverse Event Reporting System: https://vaers.hhs.gov/

[78] Figure 20: Dunlop, W. (2021, May 21). *US government database exploited by Covid-19 vaccine critics*. AFP Fact Check. https://factcheck.afp.com/us-government-database-exploited-covid-19-vaccine-critics

The report with this information spread on social media, and an article claimed the death came during ”vaccine experiments on children.” AFP Fact Check investigated:[79] the CDC told AFP the report was “completely made up,” with a professional athlete listed as the patient and a world leader’s name used for the person who reported the incident. The names did not appear in the public version of the report, and by the time it was removed from VAERS, the damage was already done.

VAERS statistics are continually exploited to push anti-vaccine claims, part of a flood of false or misleading assertions circulating online[80]and as the source of the data is an official government body, people are more susceptible to believe the information.[81]

Yet a disclaimer on the VAERS website[82] very clearly states the limitations of the data:

> “While very important in monitoring vaccine safety, VAERS reports alone cannot be used to determine if a vaccine caused or contributed to an adverse event or illness. The reports may contain information that is incomplete, inaccurate, coincidental, or unverifiable. In large part, reports to VAERS are voluntary, which means they are subject to biases. This creates specific limitations on how the data can be used scientifically. Data from VAERS reports should always be interpreted with these limitations in mind.” (VAERS[83])

This is another example of the importance of checking the methodology of a dataset and the source of the figures but not all journalists take the time to test the data, to verify the validity of the source and the methodology (Burger, 2021, MediaNumeric interview).

In a long monologue on his show in May 2021,[84] Fox News star host Carlton Tucker used VAERS data to cast doubt on the safety of vaccines during his show *Tucker Carlson Tonight*, without mentioning the VAERS disclaimer or that not all the deaths reported on the website could be attributed to the Covid-19.

There are many examples of databases that journalists have taken at face value instead of digging into the data, checking the sources and verifying for themselves.

---

[79] Dunlop, W. (2021, May 21). *US government database exploited by Covid-19 vaccine critics*. AFP Fact Check. https://factcheck.afp.com/us-government-database-exploited-covid-19-vaccine-critics
[80] Dunlop, W. (2021a, April 20). *Article misrepresents US data on deaths after vaccinations*. AFP Fact Check. https://factcheck.afp.com/article-misrepresents-us-data-deaths-after-vaccinations
[81] Wadman, M. (2021, May 26). *Antivaccine activists use a government database on side effects to scare the public*. Science. https://www.science.org/content/article/antivaccine-activists-use-government-database-side-effects-scare-public
[82] VAERS. *Data*. (n.d.). VAERS. https://vaers.hhs.gov/data.html
[83] VAERS. *Data*. (n.d.). VAERS. https://vaers.hhs.gov/data.html
[84] Wadman, M. (2021, May 26). *Antivaccine activists use a government database on side effects to scare the public*. Science. https://www.science.org/content/article/antivaccine-activists-use-government-database-side-effects-scare-public

Even scientists have unwittingly caused misinformation to flourish. On February 12, 2020, WorldPop, a research team at the University of Southampton in the United Kingdom specialised in population mapping, published a tweet about the movements of residents in Wuhan, China before the start of the coronavirus lockdown along with speculation about how Covid-19 might spread through other countries. To illustrate the tweet, the team chose a photograph that showed the global air network. The image had no connection to the spread of Covid-19.[85]

Figure 21: Screenshot of a since-removed tweet by the WorldPop Project, as recorded by AFP Factcheck.[86]

[85] AFP Australia (2020, February 18) *This map shows flight paths worldwide -- it does not show the movement of Wuhan residents.* AFP https://factcheck.afp.com/map-shows-flight-paths-worldwide-it-does-not-show-movement-wuhan-residents

[86] Screenshot of a since-removed tweet by the WorldPop Project, as recorded by AFP Factcheck. AFP Australia (2020, February 18). This map shows flight paths worldwide -- it does not show the movement of Wuhan residents. https://factcheck.afp.com/map-shows-flight-paths-worldwide-it-does-not-show-movement-wuhan-residents Archived screenshot of the tweet can be found here: https://web.archive.org/web/20200211004916/https://twitter.com/WorldPopProject/status/1225132600420917254

Although WorldPop's tweet did not itself gain much traction, the tweet was shared by a Hong Kong human rights activist from where it quickly went viral. The report was picked up by news media across the world, including UK-based newspapers Metro and The Sun; Australian newspaper The New Daily; Australian broadcaster 9 News; Australian news site News.com.au; Australian television show Sunrise; and the New Zealand-based newspaper The New Zealand Herald.[87]

A University of Southampton spokesman told AFP: "The map simply shows the global air network and the connectivity between countries using this form of travel. I think it was used as a general illustration in a tweet and somehow an incorrect story has spiralled from here." (AFP Australia, 2020, February 18).

An unrelated graphic used without explanation or qualification, combined with journalists taking the report at face value with no further enquiry, caused a map showing air transport links to create panic about the spread of Covid-19.

7.1. War in Ukraine, Conflict in Gaza, 2024 Election Year

Any major news event today is now accompanied by a growing amount of misinformation and disinformation, data analysis and data misuse. The war in Ukraine exposed the extent of Russian disinformation tactics, from peddling corruption stories about Ukraine's leadership to creating counter narratives to the massacre of civilians in Ukraine, from sowing fear among ordinary Ukrainians to attempts to sway opinion and provoke doubt in nations neighbouring and across Europe. In-depth analysis has already been carried out on this conflict by researchers affiliated with the European Digital Media Observatory and their work can be read on a special web page dedicated to the war in Ukraine.[88] Further country-specific coverage of data misuse and disinformation can be found on each of the 14 EDMO hubs that track disinformation campaigns across the 27 EU member states. Their research can be found on each of the hub's pages. [89]

Similar scenarios have been emulated to varying degrees in the conflict between the Israeli government and Hamas in Gaza[90] and in China's attempt to sway voters in Taiwan's presidential and legislative election in January 2024 (Yang, W., 2024, Jan. 5).[91]

---

[87] AFP Australia (2020, February 18) *This map shows flight paths worldwide -- it does not show the movement of Wuhan residents.* AFP https://factcheck.afp.com/map-shows-flight-paths-worldwide-it-does-not-show-movement-wuhan-residents

[88] European Digital Media Observatory (n.d.) *War in Ukraine.* https://edmo.eu/thematic-areas/war-in-ukraine/

[89] European Digital Media Observatory (n.d.) EDMO Hubs. https://edmo.eu/about-us/edmo-hubs/

[90] Hsu, S., (2023, October 14) *Analysis of cognitive warfare and information manipulation in the Israel-Hamas war 2023*, Taiwan AI Labs, https://ailabs.tw/uncategorized/analysis-of-cognitive-warfare-and-information-manipulation-in-the-israel-hamas-war-2023/

[91] Yang, W. (2024, January 5) *Q&A: Taiwan AI Labs Founder Warns of China's Generative AI Influencing Election*. VOA News https://www.voanews.com/a/q-a-taiwan-ai-labs-founder-warns-of-china-s-generative-ai-influencing-election-/7428717.html

These reports show in detail how the techniques described above are applied in real-world scenarios.

8. Artificial Intelligence

The release of ChatGPT in November 2022, and the subsequent release of other Generative AI tools such as DALL-E, Gemini and Midjourney have had such a disruptive impact on the world of storytelling and society as a whole, as well as opening up a whole new area of mis- and disinformation, that the MediaNumeric partners felt it was necessary to include a series of lectures around it as part of the online MediaNumeric Academy and address this topic in this report.

8.1. What is Artificial Intelligence and Generative AI?

IBM describes Artificial Intelligence (AI) as "technology that enables computers and machines to simulate human intelligence and problem-solving capabilities." (IBM, n.d.)[92] In order to do this, computing systems have been fed massive amounts of data, this data is processed, so that the machines can learn from the past in order to carry out very complex tasks with great proficiency, including those that require reasoning, decision-making and problem-solving.

Artificial intelligence emerged as a technology in the 1950s.[93] Over the years it has permeated all parts of life today, moving from use in science, technology and medicine to common applications that can be used by anyone, such as customer service chatbots, speech recognition, voice assistants such as Alexa, Google Assistant and Siri, household appliances such as autonomous vacuum cleaners or autonomous and self-driving vehicles.

AI is an umbrella term for this field of science of which there are a number of subsets, including deep learning, natural language processing (NLP) and large language models (LLMs) and Generative AI.

Deep learning uses multi-layered neural networks to simulate the complex decision-making power of humans (IBM; n.d.).[94]

[92] IBM - *What is AI?* (n.d.). IBM. https://www.ibm.com/topics/artificial-intelligence
[93] Copeland, B.J. (2024, March 6). *artificial intelligence*. Britannica. https://www.britannica.com/technology/artificial-intelligence
[94] IBM - *What is deep learning?* (n.d.). IBM. https://www.ibm.com/topics/deep-learning

Natural language processing (NLP) combines computational linguistics with statistical and machine learning models to enable computers and digital devices to recognise, understand and generate text and speech (IBM; n.d.).[95] It forms the basis of automatic text and audio transcription, text and voice recognition, automatic translation, and natural language generation.

Large language models (LLMs) are foundation models trained on vast amounts of data so that they are capable of understanding and generating national language and other types of content to carry out a wide range of tasks (IBM; n.d.; Maerian, 2024, Feb. 7).[96] [97]

NLP and LLMs are at the core of what is known as **generative AI,** a subset of AI that uses deep learning models to generate high-quality text, images and other content based on the data on which it was trained.[98] [99] These include chatbots such as ChatGPT, Google's Gemini or Mistral, that can process and generate an extensive range of text, from summaries, to poems, jokes, essays, computer code.

Image interfaces such as Midjourney, DALL-E, Gemini, Sora, Runway, Pika, Leonardo or Ideogram are capable of producing visuals that are limited only by the imagination of the person entering the image-generating text prompts into the software but also on the data they are trained, resulting in issues with copyright[100] as well as significant bias[101].

Audio cloning software, such as Eleven Labs, is now so advanced that it can replicate human voices that are indiscernible from the real thing, using an audio sample of just 30 seconds (Rouxel, A., 2023, MediaNumeric interview).

Generative video is lagging behind but is nevertheless making great strides. Although it is not yet publicly available, OpenAI caused a stir in February 2024 when it released a glimpse of what its new text-to-video tool Sora[102] is able to generate. Jeong Joon Park, an assistant professor of electrical engineering and computer science at University of Michigan told Scientific American that he was surprised at how fast the technology had evolved. '"I didn't expect video generators to improve this fast, and the quality of Sora completely exceeded my expectations," Park said.[103]

---

[95] IBM - *What is natural language processing (NLP)?* (n.d.). IBM. https://www.ibm.com/topics/natural-language-processing

[96] IBM - *What are large language models?* (n.d.). IBM. https://www.ibm.com/topics/large-language-models

[97] Mearian, L. (2024, February 7) *What are LLMs, and how are they used in generative AI?* Computer World https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html

[98] IBM - *What is generative AI?* (n.d.). IBM. https://research.ibm.com/blog/what-is-generative-AI

[99] McKinsey & Company. (2023, January 19) *What is generative AI?* McKinsey & Company https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

[100] https://garymarcus.substack.com/p/things-are-about-to-get-a-lot-worse

[101] https://garymarcus.substack.com/p/covert-racism-in-llms

[102] https://www.google.com/url?q=https://openai.com/sora&sa=D&source=docs&ust=1712769500743279&usg=AOvVaw3BdzxBFT8of3eERdZ720Vj

[103] Leffer, L. (2024, March 4) *Everything to Know About OpenAI's New Text-to-Video Generator, Sora.* Scientific American. https://www.scientificamerican.com/article/sora-openai-text-video-generator/

8.2. AI in the Newsroom

Every major international news organisation is collaborating in some form or another with a company that is developing tools based on artificial intelligence. These include automatic and real-time speech-to-text transcription, facial recognition to identify publicly-recognised people in photos and video, language translation, the extraction of named entities (real-world subjects such as names of people, locations, companies, topics and themes), data analysis and the automation of straightforward articles such as basic financial or sport news, among many others. These tools can process in seconds, mundane tasks that it would take journalists perhaps hours to carry out accurately.

Such examples of news collaborations in this field include the Associated Press which in 2018 teamed up with Trint to work on a video and audio transcription tool which is now widely used in major newsrooms around the world (Associated Press, 2018, May 30)[104]. Associated Press also agreed in 2023 to open part of its news database to the AI research and deployment company OpenAI to help train its system. In exchange, AP will benefit from OpenAI's expertise in the area of news generation. AP already automates some news reporting such as sports results and corporate earnings reports[105] (O'Brien, 2023, July 13). Microsoft is a major backer of OpenAI, which is behind the technology ChatGPT, DALL-E and Sora. OpenAI also signed a six-year deal with the image library Shuttershock to licence their images, videos, music and metadata in order to train OpenAI's image models (Roth, 2023, July 11)[106].

For its part, Agence France-Presse uses facial recognition software to quickly sort its extensive photo library and a tool developed in partnership with Kairntech to extract named entities from text. AFP is also extensively involved projects to develop AI-powered digital verification tools, such as InVid-WeVerify and the Vera.ai project to identify manipulated images and videos as part of its efforts in debunking mis- and disinformation circulating around the internet[107][108].

While AI tools are appearing in newsrooms around the world, editors stress that this does not change the fundamentals of journalism.

---

[104] Associated Press (2018, May 30). *AP to automate video, audio transcription with Trint.* Associated Press https://www.ap.org/media-center/press-releases/2018/ap-to-automate-video-audio-transcription-with-trint/.
[105] O'Brien, M. (2023, July 13) *ChatGPT-maker OpenAI signs deal with AP to license news stories*. Associated Press https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a
[106] Roth, E. (2023, July 11) *OpenAI's DALL-E will train on Shutterstock's library for six more years*. The Verge https://www.theverge.com/2023/7/11/23791528/openai-shutterstock-images-partnership
[107] AFP (2022, September 19) *Artificial Intelligence versus Disinformation: AFP partner in the European project vera.ai.* AFP press release https://www.afp.com/en/agency/press-releases-newsletter/artificial-intelligence-versus-disinformation-afp-partner-european-project-veraai
[108] https://www.afp.com/en/medialab

"AI is not a magic wand that can do anything and everything but, with oversight and supervision, it can eliminate time-consuming tasks and free up time so that journalists can focus on more interesting, high-value work."

- *Edouard Guihaire, Deputy Technical Editor-in-Chief (France), Agence France-Presse (2023, MediaNumeric Interview)*

In today's world when news is at everyone's fingertips, speed is essential and these tools are a game-changer in processing news and events at speed.

Some news organisations are also using AI to generate whole articles. The Executive Chairman of News Corp Australia, Michael Miller, told the WAN IFA World News Media Congress in Taipei in 2023 that News Corp was supplementing local journalism in Australia with 3,000 articles a week created by generative AI, covering topics such as local weather, fuel prices, and traffic conditions (Roper, 2023, July 7).[109]

The United Kingdom's second biggest regional news publisher, Newsquest, is using "AI-assisted reporters" to generate copy across its local titles. The group's CEO Henry Faure Walker told the Press Gazette that Newsquest had created their own AI-copywriting tool based on ChatGPT. Journalists input a press release, minutes from local planning committees or other trusted sources into the tool, set the number of words required and a suggested headline (Ponsford, 2023, Nov. 16).[110] He said this was a game-changer in terms of time that created space for reporters to spend more time out in the field. As an example, Faure Walker told the Press Gazette that the system had enabled a reporter for their northern title Hexham Courant to spend more time on a high value story when an internationally-known tree located at the famous Sycamore Gap on Hadrian's Wall near Hexham was felled by vandals:

[109] Roper, D. (2023, July 7) *Michael Miller on how NewsCorp Australia has transformed its journalism and business.* World Association of News Publishers https://wan-ifra.org/2023/07/michael-miller-on-how-newscorp-australia-has-taken-a-stand-and-transformed-its-journalism-and-business/

[110] Ponsford, D. (2023, November 16) Newsquest CEO Henry Faure Walker on bucking the trend of regional press decline., Press Gazette. https://pressgazette.co.uk/publishers/regional-newspapers/newsquest-ceo-henry-faure-walker-on-bucking-the-trend-of-regional-press-decline/

"The AI system reporter could pretty much hold the fort for the week, filling the paper, and it freed the other reporter to go out and do really good investigative stuff, videos, and get behind the story." Henry Faure Walker told the Press Gazette.

- *Ponsford, D. 2023, Nov. 16*

However, it is vital that news organisations are transparent about how stories are generated. Sports Illustrated came under the spotlight after the media company Futurism reported that the magazine had published articles generated by artificial intelligence and published them under fake author names, complete with profile pictures and mini biographies (Harrison Dupre, M. 2023, Nov. 27). Futurism said that it had found the pictures of these so-called reporters on a website that sells AI-generated headshots. Sports Illustrated refuted the report and said it was launching an internal investigation. It also removed from its website all the reporters identified by Futurism and their articles.[111]

By failing to be transparent about the origin of content and worse, actively creating a scenario which presents a fictitious person as a real journalist feeds a narrative that mainstream media is untrustworthy. "While there's nothing wrong in media companies experimenting with artificial intelligence, "the mistake is in trying to hide it, and in doing it poorly," said Tom Rosenstiel, a University of Maryland professor who teaches journalism ethics, told Associated Press. "If you want to be in the truth-telling business, which journalists claim they do, you shouldn't tell lies," Rosenstiel said. "A secret is a form of lying."" (Bauder, D. 2023, Nov. 29).[112]

This also fosters mistrust among journalists who fear that AI will be used to replace them, dumb down content and shrink newsrooms. There are diametrically opposing viewpoints within newsrooms and among storytellers more widely about the new tools that are available today. While some journalists take this evolution in their stride, there are others who are deeply distrustful of technology that they fear could, one day, replace them. Much like Hollywood screenwriters who went on strike for almost five months in 2023 to secure guarantees that studios would not replace their work with AI-generated material, many journalists are concerned that they could be writing themselves out of a job by collaborating with AI-development companies. Much like for data journalism, when there is fear, journalists are less likely to embrace the technology and learn how to use it. This, however, would be a serious handicap in navigating today's world.

[111] Harrison Dupre, M. (2023, November 27) *Sports Illustrated Published Articles by Fake, AI-Generated Writers.* Futurism. https://futurism.com/sports-illustrated-ai-generated-writers

[112] Bauder, D. (2023, November. 29) Sports Illustrated is the latest media company damaged by an AI experiment gone wrong. Associated Press https://apnews.com/article/journalists-ai-counterfeit-writers-479cc3869c0638df5bbb26d4b1e4f18f

"There is no point in developing a phobia against a technology, because this technology is here and it's not going away. Phobia can lead to a form of paranoia about the technology and a rejection. You have to learn how to use it … because it's by knowing the technology that we can avoid mistakes, that we can talk about it and report on it."

- *Edouard Guihaire, Deputy Technical Editor-in-Chief (France), Agence France-Presse (MediaNumeric Interview)*

Edouard Guihaire, Deputy Technical Editor-in-Chief at Agence France-Presse and the newsroom's representative in AFP`s working groups on artificial intelligence and innovation, said that training is key and that it was vital to bring the newsroom along with the development and implementation of new technology.

News organisations are also harnessing the power of AI to analyse archives and current production to detect bias, stereotyping and discrimination along gender and racial lines within their own reporting. Two such projects are Bias Blocker (Nguyen, K., 2023, Sept. 18)[113] and AIJO[114]: projects that were developed as part of the JournalismAI Fellowship programme, sponsored by Polis at the London School of Economics and the Google News Initiative. Revealing and recognising biases is the first step in addressing the problem and rectifying it. This can only lead to fairer and more nuanced reporting and help tackle imbalances in society as a whole.

Away from reporting, media groups are also using AI to power their own content management systems both for internal and external use. With archives creaking under the weight of the amount of content, AI is able to sort, categorise, retrieve and deliver content at speed. This is a game-changer for media organisations which are all facing challenges with financial models due to the profound shift in the way that people access and consume news content.

[113] Nguyen, K., Baris-Schlicht, I., Altiok, D., Mortada, S., Waleed, K., Taouk, M. (2023, September 18) *BiasBlocker: We asked a language model to identify racism and it tried to erase baby Hitler.* JournalismAI. https://www.journalismai.info/blog/we-asked-a-language-model-to-identify-racism-and-it-tried-to-erase-baby-hitler

[114] AFP. (2020, December 8) Using AI to automatically track bias and stereotypes in journalistic content. AFP. https://www.afp.com/en/inside-afp/using-ai-automatically-track-bias-and-stereotypes-journalistic-content

8.3. AI in Data Journalism

Artificial intelligence is fundamental in the analysis of large data sets and it is used in data journalism work across the world. It is extremely difficult and prohibitively time-consuming (and therefore expensive) for humans to extract understanding and insight from massive data sets so journalists use machine learning approaches to define a pattern in a data set that they think is newsworthy. They then use that machine learning model to find other patterns in the data that are similar to the one that they want to find. This enables reporters to cover a lot more ground in terms of the stories that they're able to find within the data.

AI was used in all the ICIJ investigations cited in section 5.2.2 and more recently, news organisations have been using artificial intelligence to examine massive amounts of satellite imagery of the conflict zones in Ukraine and Gaza as part of their coverage of news stories (Guess, 2024, Jan. 10).[115]

JournalismAI, with the support of the Google News Initiative, selects a range of different projects every year that brings together journalists and technologists from around the world. A glance at the 2024 Fellowship cohort[116] demonstrates the diversity, and the breadth of how AI is being explored within newsrooms (Sivadas, 2024).

1. AURA: Advanced Understanding and Research Assistant: A conversational AI platform designed to help journalists navigate and extract stories from complex unstructured data, for example, in science and technology. By leveraging generative AI, Aura offers instant context and investigative leads, enhancing the depth and quality of journalistic content.

2. IntelliNewsComparer: GenAI-based document comparison tool: The tool will employ machine learning to semantically compare text documents in English, Finnish, and Tagalog. Utilising pre-trained LLMs will enable swift, precise vector searches with a user-friendly interface for querying document similarities. Adaptable to multiple documents and LLM sources, it can expedite the examination of documents in investigative newsroom projects.

3. CheckMate: A web app for real-time fact-checking on live or recorded video and audio broadcasts to enable journalists to easily identify and debunk false claims and enable newsrooms to actively work to prevent the spread of misinformation over a significant election year.

4. Real Estate Alerter: A tool to harness anomaly detection methods and LLMs to uncover hidden news stories within real estate data. The goal is to create an alert system, providing

[115] Guess, R. (2024, January 10) How AI Could Act as Boost for Investigative Journalism. VOA New. https://www.voanews.com/a/how-ai-could-act-as-boost-for-investigative-journalism/7434364.html
[116] Sivadas, L. (2024, March 11) *Welcoming the 2024 cohort of JournalismAI Fellows*, JournalismAI. https://www.journalismai.info/blog/welcoming-the-2024-cohort-of-journalismai-fellows

reporters with a head start on newsworthy stories, using AI to gain an edge in breaking news and investigative journalism.

5. StyleCheck: An AI-driven tool to check adherence of articles with the newsroom's style guide. Integrated with a cleanup tool for article copy, it would provide suggestions that can be overwritten. The goal is to apply guidelines more consistently, thus improving the quality and credibility of news.

6. Data Robot Aide: An open framework for data-driven and LLM-powered AI chatbots to help newsrooms create chatbots using different structured datasets, like elections, census, crime, healthcare, etc. These chatbots will speed up storytelling and help find initial story ideas.

New AI-powered tools are released regularly. One such tool is Google Pinpoint which allows users to upload and search hundreds of thousands of documents, images, emails, hand-written notes, pdf files and audio files to search for specific words, phrases, locations, organisations, or people. It can transform similarly structured tables into sortable spreadsheets. A single Pinpoint collection can contain up to 200,000 documents.[117]

Natalia Żaba is an investigative journalist and teaching fellow at the Google News Initiative. She says that this tool has made a massive difference in the types of stories that journalists are able to take on. With the ability to sift massive amounts of data, it is possible in the early stages of reporting to figure out whether there is enough material to move ahead with an investigation. It saves a crucial amount of time.

"There is no way that an editor-in-chief can dedicate a reporter to read 60,000 court documents. That is just impossible. Even if the newsroom has the financial resources to dedicate a team to work on a case for one year, how much of this work do you want done by your people, just reading one page after another?" Żaba said (Żaba, 2023, MediaNumeric interview).

"Today, it seems a little bit pointless. Reporters should be dedicated to the thinking and analyzing phase, working with sources, which is really the core of our job. AI does the boring job that we normally don't want to be doing."

Responding to fears by journalists about how AI is changing the media landscape, she said that journalists were right to be cautious and she advocates for regulation but she also said that this time of change provides an opportunity to do things better.

[117] https://journaliststudio.google.com/pinpoint/about/

"It's not about technology really, it's about us, humans, and how we use it, to be able to use the tech for the good things, to elevate the journalism, to bring the stories that we really want to see, to start thinking from another level. This is a big opportunity for all of us."

- *Natalie Żaba, Investigative journalist and Google News Initiative Teaching Fellow (MediaNumeric interview)*

8.4. AI in Misinformation and Disinformation

The evolution of generative AI in images, audio and video is advancing at breakneck speed allowing people peddling misinformation to pump out original, but false content that may appear totally plausible. Artificial intelligence has been widely used by fact-checkers for years to debunk manipulated content, either by using reverse image tools to check whether an identical or similar image has been published previously, or by analysing the image or audio itself to identify tell-tale markers left behind in the manipulation process.

With the arrival of generative AI, fact-checkers, journalists and ordinary citizens now have to contend with images, text and audio that look entirely like the real thing but that are 100 percent artificial. This "synthetic" content leaves behind no manipulation markers that journalists can use to identify the content as reused.

Examples of audio cloning are appearing at an alarming rate, from a recording that claims to show the Mayor of London Sadiq Khan making inflammatory comments about Armistice Day in the United Kingdom to a robocall of US President Joe Biden in January 2024 telling New Hampshire residents not to vote in the Democratic primary. It is becoming increasingly difficult to know what is real or not. As detection tools are struggling to keep up, fact-checking organisations can only rate information as "No evidence that it is true"[118] (Thompson, 2023, Nov. 10), which leaves open the door of doubt and enables disinformation to flourish.

There is a growing number of groups using AI to create tools to identify synthetic content but experts say that they are always one step behind.

> "Machine learning is really good at telling you about something it's seen before, but it's not so good about reasoning about things it hasn't seen," Patrick Traynor, a University of Florida

[118] Thompson, T. (2023, November 10) *No evidence clip of Sadiq Khan supposedly calling for 'Remembrance weekend' to be postponed is genuine*. Full Fact. https://fullfact.org/news/khan-audio-palestinian-remembrance/

> professor who specialises in computer science and telephone networks, told the NBC news network.[119] (Collier, K. & Cui, J., 2024, Feb. 4)

With accuracy and veracity absolutely essential to all reputable media groups, newsrooms are teaming up with tech companies and researchers to develop tools to identify these so-called deep fakes. AFP is participating in a European-wide project, funded by the European Union, called Vera.ai[120] to do just that. Denis Teyssou, who is leading the research and testing within AFP says that it is a long process to create a reliable system. He says an additional challenge is the method used to build the tool is published in scientific papers. Open data and transparency is important in the fact-checking world, however the publication of papers gives companies that create deep fake technology a way to counter detection technology and improve their own cloning software. (Teyssou, 2023, MediaNumeric interview).

With the new tools they are developing, the Vera.ai consortium is restricting the tools' use to journalists, fact-checkers and researchers who work on disinformation "to prevent people who want to make better fakes from using our tool to see whether or not it picks up their fake," Teyssou said (Teyssou, 2023, MediaNumeric interview).

Neil Zhang, a machine learning researcher at the University of Rochester, USA, told NBC that while these projects are vital in stemming the flow of misinformation and disinformation, they face an uphill struggle.

> “There’s a huge disparity in funding between companies racing to make passable deep fakes versus those trying to detect them,” Zhang said, cited by NBC. “It’s hard to get funding for detection, very easy to get funding for large-language models and generative AI.” (Collier, K. & Cui, J., 2024, Feb. 4)[121]

8.5. Using Content Authenticity to counter misinformation

One of the approaches to ensuring the authenticity and provenance of digital content is to use cryptographic asset hashing. This can be employed to furnish verifiable, tamper-evident signatures indicating that neither the image nor its associated metadata has undergone undisclosed alterations. The use of this approach is advocated through the Coalition for Content Provenance and Authenticity (C2PA).[122] By formulating technical standards and protocols for embedding metadata into digital content, such as images and videos, C2PA lays the groundwork for verifying the genuineness and origin of such content. C2PA unifies the efforts of the Adobe-led Content

[119] Collier, K. & Cui, J. (2024, February 4) *Why AI-generated audio is so hard to detect.* NBC https://www.nbcnews.com/tech/misinformation/ai-generated-audio-detect-tool-model-rcna136634
[120] Vera.ai https://www.veraai.eu/home
[121] Collier, K. & Cui, J. (2024, February 4) *Why AI-generated audio is so hard to detect.* NBC https://www.nbcnews.com/tech/misinformation/ai-generated-audio-detect-tool-model-rcna136634
[122] https://c2pa.org/

Authenticity Initiative[123] which focuses on systems to provide context and history for digital media, and Project Origin,[124] a Microsoft- and BBC-led initiative that tackles disinformation in the digital news ecosystem.

This transparency fosters a clearer understanding of content creation and alteration processes, empowering journalists, fact-checkers, non-governmental organisations (NGOs), and news consumers to discern credibility more effectively. Additionally, C2PA's support for the development of verification tools enables media organisations to authenticate user-generated content and identify instances of misinformation or manipulation. Through promoting these standards and tools, C2PA nurtures trust within the news ecosystem, reinforcing the reliance on verifiable information and mitigating the dissemination of falsehoods. C2PA's endeavours are pivotal in promoting transparency, trust, and authenticity in the ever-evolving landscape of digital media.

The growing update of this approach is demonstrated by the announcement by OpenAI to include C2PA to images generated with ChatGPT on the web and their API serving the DALL·E 3 model metadata. From the press release:

> "People can use sites like Content Credentials Verify to check if an image was generated by the underlying DALL·E 3 model through OpenAI's tools. This should indicate the image was generated through our API or ChatGPT unless the metadata has been removed."[125]

All major camera manufacturers such as Nikon, Canon, Sony and Leica are working to integrate C2PA technology into their new models and many of these manufacturers are working hand-in-hand with major image generators, such as Agence France-Presse, the Associated Press, Reuters and Getty Images, to build the framework, test and ultimately implement this technology. The news generators see it as fundamental to fighting the erosion of trust in the news media (Dulai, 2023)[126].

8.6. Challenges in Using AI in Newsrooms

One of the biggest challenges in using artificial intelligence, not just in newsrooms but across the board, is who owns the data and the computational process used to sort and analyse the data. Many artificial intelligence initiatives that began with a public interest mandate have changed over the years as major companies such as Microsoft or Google have taken increasing, or even controlling stakes in their activities or have even bought up start-ups outright. As the technology passes into the hands of private companies, it becomes difficult to know how the LLMs were put together, with what content and what kind of algorithms. Algorithm bias can totally change the

[123] https://contentauthenticity.org/
[124] https://www.originproject.info/
[125] https://help.openai.com/en/articles/8912793-c2pa-in-dall-e-3
[126] Dulai (2023, November 22) *Sony completes second round of AP testing of C2PA in-camera authenticity technology* https://www.dpreview.com/news/9855773515/sony-associated-press-test-in-camera-authenticity-technology

type of results the AI provides. Even though he had already participated in the creation of multiple AI tools to debunk misinformation, Denis Teyssou said that the Vera.ai debunking project went back to basics so that they could be sure that the software would be built on solid foundations. With the previous tools, "it was the machine, the black box of AI, which would decide what was real or false based on criteria that were not necessarily robust because they weren't known and because they could not be explained. We are now looking for things that are much more precise." (Teyssou, 2023, MediaNumeric interview)

Alexandre Rouxel, Senior Project Manager specialised in Data and AI at the European Broadcasting Union (EBU), said the EBU decided to build their own LLM where they designed the architecture, developed, trained and tested the models. He says that it is difficult to use the system to identify disinformation as it can depend on the context.

> *"A number can be true, depending on the context so you have to understand the context in which the number was presented. That can be difficult for AI. A journalist can understand humour which can refer to things that happened long ago. This is why it is difficult to make a firm decision. There are not many AI tools that can do this because it depends too much on the context."*
>
> - *Alexandre Rouxel, Senior Project Manager, European Broadcasting Union (MediaNumeric interview)*

The EBU is the world's leading alliance of public service media. It has 113 member organisations in 56 countries with an additional 31 associates in Asia, Africa, Australasia and the Americas. Its members operate nearly 2,000 television, radio and online channels and services and together can reach an audience of more than one billion people around the world broadcasting in more than 150 languages (EBU, undated)[127]. They have many different missions but at its core, they work to foster better connections between its members and provide a framework with the aim of telling meaningful stories. As such, the EBU remains at the forefront of technology. To address the difficulty of AI and its ability to identify mis- and disinformation, the EBU created tools to analyse language, including language often used in the context of disinformation, so that they can attribute articles with a "reliability score" and thus warn journalists that certain statements should be carefully checked.

[127] https://www.ebu.ch/about

While large news organisations may have the financing and the technical expertise, how do smaller newsrooms manage this challenge in the years to come? Also, the lion's share of AI technology is written in English, making the technology harder to access in non English-speaking countries.

The role of artificial intelligence within newsrooms, data journalism and storytelling is extensive and expanding extremely fast. It is beyond the scope of this report to examine this issue in fine detail. Professor Charlie Beckett, Director of Polis and the Polis/London School of Economics JournalismAI project, and lead researcher Mira Yaseen, published in 2023 an excellent in-depth report on the topic.[128]

9. The Role of National Audiovisual Archives

The collaboration between industry, academic and the archive in MediaNumeric has provided a rich source of information and areas of exploration. National archives hold the history of a country, or, at least part of it. However, working with these archives present considerable challenges in and of themselves. Much of the older content has no metadata at all. It can be very difficult, even impossible, for researchers to find what they are looking for when much of the content is only partially labelled, or not labelled at all. It is an uphill struggle just to know what is there. Even when content is labelled, the information may be sketchy and imprinted with the social norms and societal attitudes of the time, some of which could potentially be seen as racist or sexist by today's standards. Yet, it is crucial that this effort is undertaken. Archives are an undervalued mine of information for news and data journalists. They can also be used to debunk misconceptions, cliches and prejudices that can fuel modern-day misinformation, as can be seen in the examples below.

9.1. Managing and Sharing an Ever-Increasing Database

The amount of data that national archives must manage is exploding. Thankfully, there are now processes available and in place to automatically add extensive and rich metadata to new content being generated today. However, there is a significant challenge in annotating older archives so that they can be made visible, be found and be used. National archives are faced with a technological transformation. Data and artificial intelligence is key to this transformation as the sheer volume of content to categorise is impossible for humans to manage.

The EBU is developing its own facial recognition tool to scour the archives which it will make available for its members. EBU's Senior Project Manager Alexandre Rouxel says, although there are many facial recognition tools available on the market, it was important for EBU to develop their

[128] Beckett, C., Yaseen, M. (2023, September 20) *Generating Change. A global survey of what news organisations are doing with AI.* Polis. [downloadable pdf] https://www.journalismai.info/research/2023-generating-change

own tool so that they could train it to recognise local or national celebrities which may not be well known internationally and therefore indiscernible to off-the-shelf software. By developing their own tool, the EBU can be much more precise and accurate and therefore make much better use of the archives. In controlling the LLM, the EBU can also ensure that they comply with General Data Protection Regulation (GDPR) and thereby only record the names of people who fall within what is considered the public domain.

Speech-to-text tools also enable national archives to annotate huge amounts of content, which Rouxel described as a "game-changer" for archive repositories.

> "It's impossible to multiply by 10 or 100 the number of archivists. The fact that artificial intelligence is arriving with mature technology, may change the way that they work. The more mundane and less interesting tasks could be taken over by artificial intelligence, which means that archivists could take on roles that are more creative." (Rouxel, A., 2023, MediaNumeric interview)

Rouxel added that the work being done at the European Broadcasting Union has benefits far beyond just the EBU. Effective and reliable data enables EBU members to share and exchange content and data in ways that had simply not been possible before and thereby strengthen the resilience of the media grouping.

9.2. Working with National Archives

As with data journalism, graphics, videographics and fact-checking, the Covid pandemic was a turning point for audiovisual archives. With almost no ability to film new content, television producers and journalists turned to the archives to fill their news and television programmes.

Hélène Rauby Matta, a Business and Training Development Manager at the EBU Academy, remembers how the relationship between the news desk and archivists, or media asset managers, changed during that period.

> "Quickly we had a request from our members who found themselves working from home, including journalists and reporters and producers and who had little material to work with and who suddenly started, I wouldn't say rediscovering their archives, but certainly paying more attention. Shooting was really difficult for several months. So they started going back to their colleagues to establish that connection and started maybe talking again to the archivists," Rauby Matta said. (Rauby Matta, 2021, MediaNumeric interview).

Rauby Matta describes these national audiovisual archives as an "untapped resource and capital" however she said that many journalists require training in order to be able to search archives efficiently and that they may also need guidance to know how to extract stories from an archival

database. As a development manager for the EBU Academy, she says that they are training an increasing number of journalists in AI, data and big data tools.

> *"We train dozens of journalists on building constructive stories, and that is also offering multifaceted approaches to issues and storytelling that is not just based on the one event, but may be going back in the past. I think this has helped redefine journalism a bit for some of our members. I think this has had a positive impact on reconsidering how you can use archives to actually feed and enrich the stories."*
>
> - *Hélène Rauby Matta, Business and Training Development Manager, European Broadcasting Union Academy (MediaNumeric interview)*

Rauby Matta continued: "It's the same approach in a way when, as a public service media, you try to debunk fake news or you try to educate your team. It's also about substantiating news stories with past elements with historical facts for which again you may have to dig into your archives and do a bit more of research." (Rauby Matta, 2021, MediaNumeric interview).

Willemien Sanders is a media scholar affiliated with Utrecht University and the national Sound and Vision audiovisual archives (NISV) in the Netherlands who does just that. Sanders is specialised in telling stories with data and her research has shown that national archives contain very powerful stories, originated from both contemporary content and using archives to change the narrative of modern-day viewpoints.

By analysing airtime for various politicians over a given timeframe in recent years, Sanders was able to identify that certain political parties appeared comparatively infrequently in the media which meant that their voice and their ideas were not heard.

> "I used to refer to it [eds: NISV] as the history of the public debate in the Netherlands, but I increasingly realised that it's a very limited public debate because a lot of groups were never included or never given a chair in that debate, they were never given an opportunity to participate in that debate. That is something that I think we should realise is problematic," Sanders said (2023).

Sanders believes that there are benefits to bringing members of under-represented groups in society into the archives so that they can provide a different light on what is available in the

archives, how that content is portrayed. They would also be able to more accurately identify what is missing from the archives. She stressed that it was important to involve members of immigrant communities in the process. People who immigrated decades ago may have spoken a form of language or dialect with which second and third generation immigrants are no longer familiar. Without the knowledge of these spoken traditions and stories, we risk losing parts of our heritage.

Yvonne Ng, a digital archivist at the human rights advocacy NGO Witness, concurred. "You don't think of data collection in a vacuum. Data is information about people and, especially in the context that I work in, we're very conscious of that and we always like to work in partnership with local organizations or people who are directly impacted by the data. We're not really interested in data collection for the sake of data collection but always for a purpose." (Ng, 2021, MediaNumeric interview).

In an interview with the MediaNumeric team, and during the final event of the MediaNumeric programme, Sanders explained that when certain groups of people were always represented in a specific light, the constant reuse of those images reinforces and perpetuates a trope of that group of people, be they rich, poor, or from specific ethnic, religious or social backgrounds. While there was a danger in fixing people into one non-evolving identity, Sanders said that alert documentarians could use the archives to change today's stereotypes. As an example, Sanders said that people of Moroccan, Turkish or Yugoslav backgrounds are today identified by some in the Netherlands as passive and not integrated into society. They are viewed by some as a problem. However, through her investigation into the archives, Sanders has found footage of these emigrants as active and engaged members of Dutch society.

> "This is an image or a representation that we don't often see nowadays. So by showing and really searching what was there before and diving into the archive … there's this alternative representation that sheds another light on how we look back. In the very contemporary discourse on foreign workers or foreign people in the Netherlands, it's really useful to have an alternative representation that also shows: 'Hey, these people are here because we asked them to come over and we needed them to do the work that we didn't want to do and they're not just passive or unwilling or whatever.'" Willemien Sanders said (2023, MediaNumeric interview).

A similar example comes from the National Audiovisual Institute in France. In the midst of the #MeToo movement against sexual violence against women, a data journalist at INA located in 2019 an excerpt from a high-brow literary talk show on Antenne 2 that dated from 1990.[129] In the clip, a well-known and well-respected author Gabriel Matzneff explains his preference for insatiable sex with underage girls and boys, some as young as 14. Only one of the other guests, Denise Bombardier, challenges Matzneff on his behaviour and calls it out for the crime that it is: paedophilia. The way in which Denise Bombardier was pilloried for her criticism of Matzneff goes

[129] Pivot, B. (2019, December 19) 1990 : *Gabriel Matzneff face à Denise Bombardier dans "Apostrophes"*, INA Archive. https://www.youtube.com/watch?v=H0LQiv7x4xs

some way to explaining how incest and sexual abuse among France's literary and film elite was tolerated and covered up for such a long time, including up until 2023 and accusations of sexual abuse against France's film icon Gérard Depardieu.[130]

9.3. Challenges in National Archives

An immediate challenge is bringing the archives into the core of the newsroom and reporting. Rauby Matta says that many media asset managers work in small teams and can feel isolated both from the mainstream reporting teams but also from the wider archiving community. Some archivists benefit from personal connections with others but in many organisations there is no structured process of communication so this valuable resource can remain largely untapped (Rauby Matta, 2021, MediaNumeric interview). She said there would be considerable benefit to bringing archivists to the centre of the newsroom.

> "There are still a lot of places where archivists are completely disconnected, even physically from the place where stories are produced and it's not just digital. I think sometimes just having someone sitting there helps, as a liaison office or something. You can't have everyone, everywhere at the same time but for me, it's more about connecting, but connecting in a structured corporate way.
>
> "There is this need for dialogue and to show the opportunities that can be coming from this collaboration, or also when it comes to storytelling, not just working with planning new stories." (Rauby Matta, 2021, MediaNumeric interview)

As for data journalism, Rauby Matta said the solution lay in reaching the team managers and not just training the news and beat reporters.

> "We feel sometimes that we are really not reaching out to those managers who could make the difference. It's more going bottom up when sometimes we need to go top down." (Rauby Matta, 2021, MediaNumeric interview)

Much of the audiovisual content held by national audio/visual archives is under strict copyright and it may not open to the general public. These archival institutions may not hold the rights themselves as they provide a hosting service for television channels who may ultimately retain control over their content. Some archive institutions are making more content available online but in many countries access is restricted to authorised viewers such as academics and researchers but this vastly reduces the number of people who can access this shared history of a country. In certain

[130] Le Parisien (2024, March 5) *Affaire Depardieu : l'acteur visé par une nouvelle enquête pour agression sexuelle, après la plainte d'une décoratrice*. Le Parisien. https://www.leparisien.fr/faits-divers/affaire-depardieu-lacteur-vise-par-une-nouvelle-enquete-pour-agression-sexuelle-apres-la-plainte-dune-decoratrice-05-03-2024-IXHZE4SETFFOXGIHNGDNDVHA7E.php

cases the copyright extends not just to the content itself but all of its associated metadata so even those who are granted access need to create an additional layer of independent aggregated metadata over the top so that they are then legally allowed to use the information.

National archives face exactly the same questions as those who create databases, LLMs, AI or debunking tools. The archives are costly to create and to maintain and the data has a financial value. The reuse and sale of archive material is an important revenue source for television channels and radio stations. However this does mean that only a select few are able to intervene on the content.

Sanders says that this conundrum merits a wider nationwide debate about the role of a public national broadcaster in Dutch society. "Increasingly, there's a tension between the need for reliable information and how money is made through data manipulation. That's actually a very big question … and the role of public media and what we want it to be, to do and how much we will spend on it." (Sanders, 2023, MediaNumeric interview)

In order to maximise uptake and reuse of archives, Martin Bouda, an archivist at Czechia Archives in the Czech Republic, suggested that archives should make available to journalists an interface which is simpler to navigate than the programme used by archivists which can sometimes offer a complex search environment (Bouda, 2021, MediaNumeric interview).

Another challenge is how to add metadata to archive images. While reviewing children's programmes of yesteryear, researchers came across a children's show that contained an offensive racial slur against Black people. A debate ensued as to whether the N-word should be included in the metadata. Some said yes, arguing that it was a quote, others felt that the slur should be described but not directly quoted due to the highly offensive nature of the word. Sanders said that it was important, when adding metadata, to look forward to the generations to come, to consider what type of terminology they might use in their search.

> "It's a real challenge how to make the metadata sustainable for the future, so that in a few years, those people who are still young now will probably use different terms to search for material. How can you make sure they find it?" (Sanders, 2023, MediaNumeric interview)

Sanders also stressed that it was vital that European archives come together to agree on a common standard for metadata tagging to ensure better collaboration between the different institutions. However, she highlighted differences in attitudes today towards societal issues that could lead to different ways of describing content within the archives. For example, LGBTQ+ rights are considered differently across the European Union and so the metadata describing the content will be based on different principles.

> "It's naive to think that we can create some sort of objective way to describe certain phenomena or groups or people because the terms that you use already to describe certain

> phenomena already have meaning and create meaning and assign meaning. … So really the question is, can we, or at least people in the consortium across Europe, can we agree on the approach? What do we think is the function of an archive?" (Sanders, 2023, MediaNumeric interview)

Ng said finding this common ground is extremely challenging. She said that there was a perception that there was objectivity in the science of archiving but that value judgments were being made every day during the collection process so it was important to recognise and admit the limits of the process. "I think all collections are biassed, that's like it has a negative connotation that term but you know all collecting is done with a certain perspective. There is no neutral position so it's more about being transparent." (Ng, 2021, MediaNumeric interview).

Sanders concurs:

> *"Despite all the technological developments, basically we still use a lot of human beings to develop those technologies and also to do a lot of the work. Humans with all their flaws and shortcomings and preferences and ideas and everything, so it will never be perfect. … It should be a tool, but not an end in itself."*
>
> - *Willemien Sanders, Media scholar, Utrecht University and Sound and Vision (MediaNumeric interview)*

Sanders said that it would be beneficial to create a consortium across Europe to debate these issues, to identify the function of an archive and decide on a common approach. The European-funded project TEMS aims to do just that. TEMS - or Trusted European Media data Space - brings together 43 organisations from across 14 countries in the European Union with the aim of redefining the way that the media sector shares and extracts value from data[131].

10. European Policy

On March 13th, the European parliament adopted the EU Artificial Intelligence Act, a landmark new law that the European Commission describes as the world's first-ever legal framework on AI. It aims to "foster trustworthy AI in Europe and beyond, by ensuring that AI systems respect

[131] https://tems-dataspace.eu/

fundamental rights, safety, and ethical principles and by addressing risks of very powerful and impactful AI models."[132] (European Commission, 2023)

The AI Act defines four levels of risk system for machine-learning systems according to the potential threat they pose to society: minimal risk, limited risk, high risk and unacceptable risk. Any system considered to pose an unacceptable risk will be banned and those in the high risk category, which will include, among others, AI technology used in critical infrastructures, employment, private and public services, educational and vocational training, law enforcement, border controls and justice, will be subject to strict controls before they can be implemented. Some may even require judicial approval. Those in the limited and minimal or no risk categories will have less oversight, though transparency about the AI systems remains important.

Although the Act has been approved by the European Parliament, member states must still decide which regulator is best placed to oversee compliance. While the bans on prohibited practices will come into force in November 2024. General-purpose AI rules will apply from May 2025 and it will take three years before high-risk systems will have to comply.

Legislative bodies move at much slower speed than today's technology but the European Commission stresses that the AI Act can evolve if needed.

> "As AI is a fast-evolving technology, the proposal has a future-proof approach, allowing rules to adapt to technological change. AI applications should remain trustworthy even after they have been placed on the market. This requires ongoing quality and risk management by providers."[133] (European Commission, 2023)

Reactions to the AI Act have been mixed. The tech sector fears that the regulation will stifle innovation within the European Union and swamp small start-ups in compliance paperwork, putting them at a considerable disadvantage to competitors in the United States, China or the United Kingdom (Davies, P., 2023, Dec, 15)[134].

In July 2023, the European Commission asked all tech platforms to sign up to a voluntary Code of Practice on Disinformation under which they commit to reduce the spread of disinformation and to label content generated by artificial intelligence. Meta, Google, YouTube, TikTok and LinkedIn signed up to the code but X, formerly Twitter, did not. Disinformation has increased significantly on the platform since it was taken over by tech entrepreneur Elon Musk.

---

[132] European Commission, (2023, undefined) Shaping Europe's digital future: AI Act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#:~:text=The%20AI%20Act%20is%20the,play%20a%20leading%20role%20globally.&text=The%20AI%20Act%20aims%20to,regarding%20specific%20uses%20of%20 AI

[133] European Commission, (2023, undefined) Shaping Europe's digital future: AI Act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#:~:text=The%20AI%20Act%20is%20the,play%20a%20leading%20role%20globally.&text=The%20AI%20Act%20aims%20to,regarding%20specific%20uses%20of%20AI

[134] Davies, P., (2023, December 15) *'Potentially disastrous' for innovation: Tech sector reacts to the EU AI Act saying it goes too far.* Euronews. https://www.euronews.com/next/2023/12/15/potentially-disastrous-for-innovation-tech-sector-says-eu-ai-act-goes-too-far

Six months after the launch of the Code, the European Commission Vice-President for Values and Transparency Vera Jourova hailed the platforms' initial steps but warned them that Europe expected further steps to mitigate Russian disinformation about Ukraine and a potential campaign to destabilise the European elections in June 2024 (Jourova, 2023, Sept. 20)[135].

Aside from applying pressure on online platforms to mitigate disinformation, the European Commission has funded numerous pan-European projects designed to encourage collaboration across countries in the fields of fact-checking, data journalism and analysis and artificial intelligence. MediaNumeric is one such example, but there are many others including the European Digital Media Observatory (EDMO), the European Newsroom, Vera.ai and the European Fact-checking Standards Network (EFCSN).

However, to really fight the rampant spread of disinformation over the long term, authorities need to rethink education so that people are taught from an early age the foundations of critical thinking and how to work with data.

> "We need to rethink educational systems at all levels, from curricula to teacher training," Peter Addo, Lead Data Scientist in charge of AI strategy at the French Development Agency (AFD), told an AI-focussed conference in Paris at the end of January . "We need to mainstream data and AI literacy, and develop critical thinking skills so that we question everything we see online." (Agence Française de Développement, 2023, Feb. 21)

11. Conclusion

Data-driven storytelling, digital verification and artificial intelligence are part of today's media landscape, both for producers of content and also for content consumers. Yet these fields require new skills in order to use them correctly.

In the face of the 'firehouse of information', journalists who have not yet mastered these tools run the risk of adding to the confusion, misinformation and disinformation that can spread through social media platforms and even, at times, via media organisations themselves.

In using a badly-designed visualisation, misinterpreting data or failing to challenge information provided by sources, journalists can inadvertently encourage or accentuate false beliefs that can be hard to dismantle. The manipulation of information and the speed at which it is now able to

[135] Jourova, V. (2023, September 26) *Press statement of Vice-President Jourova on the meeting with the Code of Practice on Disinformation Signatories. European Commission.* https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4645

circulate across the globe has sharpened societal divides and polarised political discourse leading to real world action as witnessed in the storming of Capitol Hill in January 2021.

Journalists who have developed these techniques are able to cut through the noise and ensure that they deliver reliable, accurate and truthful information despite the time pressure of today's 24-hour, up-to-the-minute news flow.

Yet knowledge of how to work with data, artificial intelligence and digital investigation tools efficiently and responsibly are still lacking from many newsrooms. Working with spreadsheets and statistics is off-putting to many people who have followed a humanities route. Data-driven stories are still considered a niche area and many journalists do not realise that digital verification is a separate skill set beyond simply checking facts.

Working with data and digital verification brings a huge sense of purpose to those who excel in these fields. Their work can have hard-hitting impact and lead to significant and long lasting change in society.

Training is essential to encourage more story-tellers to adopt data and digital verification skills, to foster cooperation between media to broaden the depth of stories that they are able to uncover and encourage public institutions to open up their databases to public examination. It also requires a shift in attitudes among news editors and in newsrooms as a whole.

These skills are not just important for storytellers. The data and information overload means that it is more urgent than ever to ensure that readers are able to decode information, to look at content critically and come to their own conclusions.

The MediaNumeric consortium was created to address this need.. After honing best practices over three on-site training sessions, it has developed the MediaNumeric Academy (www.medianumericacademy.eu)[136] so that other students, storytellers and the public at large are able to develop their own expertise and hopefully find the stories of the future.

[136] MediaNumeric Academy. https://www.medianumericacademy.eu/

12. References

Interviews

Interviews carried out between March and May 2021 and in November 2023 by MediaNumeric consortium members:

1. Abellan, Andrea, European Journalism Centre, Data coordinator (NL)
2. Balcerzak, Bartłomiej, Polish-Japanese Academy of Information Technology, PhD in computer science with a major in natural language processing (PL)
3. Bazán Gil, Virginia, RTVE; Member of the Stakeholder Board (ES)
4. Bastien, Karen, WeDoData, Co-founder; Member of the Stakeholder Board (FR)
5. Biel, Beata, Director of Premium Content, Discovery / TVN; Journalist, academic teacher (PL)
6. Bommenel, Alain, Agence France-Presse, Head of infographics and innovation (FR)
7. Bouda, Martin, Archivist, Czechia Archives; Česká Televisie; Member of the Stakeholder Board (CZ)
8. Broch, Louise, DR, Archivist and Researcher; Member of the Stakeholder Board (DK)
9. Bruins, Jochem, NOS, TV journalist, Jeugdjournal, news program NPO (NL)
10. Burger, Peter, nieuwscheckers.nl; Leiden University, Founder; Associate professor in the Department of Media Studies (NL)
11. Cairo, Alberto, University of Miami, Knight Chair in Visual Journalism at the School of Communications (USA)
12. Claessens, Tom, data journalist, Follow the money (NL)
13. Herve, Nicolas, Institut National de l'Audiovisuel, Senior researcher (FR)
14. Cook, Lindsey, The New York Times, Senior Editor, Digital Storytelling and Training (USA)
15. Cunliffe-Jones, Peter, University of Westminster; International Fact-Checking Network (IFCN); Africa Check; Course director; senior advisor to IFCN; founder of Africa Check (UK)
16. Dam, Yordi, Local Focus, co-founder (NL)
17. Desponds, Anna, freelance curator, expert; Adam Mickiewicz Institute, Warsaw, Dwutygodnik.com, Warsaw, Creatives' Catalysts agency, Berlin, Total Immersion Foundation, Warsaw (PL)
18. Dudfield, Andy, Full Fact, Head of automated fact-checking (UK)
19. Guihaire, Edouard, Agence France-Presse, Deputy Technical Editor-in-Chief (France)
20. Hoffman, Dirk, DAIN Studios, Co-founder and CEO (DE)
21. Holan, Angie Drobnic, PolitiFact, Editor-In-Chief (USA)
22. Jehel, Sophie, University of Paris 8 Saint-Denis, Lecturer; Member of the Stakeholder Board (FR)
23. Kaliszewska, Joanna, FINA, head of digital repository team (PL)
24. Kessler, Glenn, The Washington Post, Editor and Chief Writer of The Fact Checker (USA)
25. Kiely, Eugene, FactCheck.org, Director (USA)

26. Kraus, Daniela, Presseclub Concordia, Secretary General, Journalist; Member of the Stakeholder Board (AT)
27. Léchenet, Alexandre, Independent, La Gazette des Communes, Data journalist; Member of the Stakeholder Board (FR)
28. Linden, Johan, SVT, Project Manager Future of News (SE)
29. Lombion, Cédric, Open Knowledge Foundation & School of Data; Member of the Stakeholder Board (FR)
30. Malfatto, Simon, Agence France-Presse, data journalist, Deputy Head of infographics and innovation (FR)
31. Meeuwenoord, Mark, Avans University of applied sciences, research fellow, media artist, lecturer (NL)
32. Nack, Frank, INDE Lab, University of Amsterdam, Associated Professor (NL)
33. Ng, Yvonne, Witness, Audiovisual Archivist (USA)
34. Nicholson, Sophie, Agence France-Presse, Deputy Head of Digital verification (FR)
35. Odijk, Daan, RTL, Lead Data Scientist with a PhD in Information Retrieval (NL)
36. Orsek, Baybars, International Fact-Checking Network, Director (USA)
37. Parasie, Sylvain, Sciences Po, médialab; Professor of sociology; Member of the Stakeholder Board (FR)
38. Ptaszek, Grzegorz, AGH University, PhD in social communication and media studies and Psychology Professor (PL)
39. Rauby Matta, Hélène, EBU Academy; Member of the Stakeholder Board (CH)
40. Rendina, Marco, Istituto Luce - Cinecittà, International projects coordinator; Member of the Stakeholder Board (IT)
41. Reynolds, Sally, Audiovisual technologies, informatics & telecommunications ATiT, Director and co-founder (BE)
42. Rippon, Peter, BBC Online Archive, Editor; Member of the Stakeholder Board (UK)
43. Rogers, Richard, University of Amsterdam, Professor of New Media and Digital Culture (NL)
44. Rouxel, Alexandre, European Broadcasting Union, Senior Project Manager, Data and AI (CH)
45. Rzeczkowski, Grzegorz, Polityka Weekly, Chief investigative reporter (PL)
46. Sanders, Willemien, Utrecht University, Media scholar, data stories expert and affiliated researcher at the Institute for Cultural Inquiry and at Sound and Vision (NL)
47. Schuth, Anne, DPG media, Data Science lead (NL)
48. Tarkowski, Alek, Open Future Foundation, Strategic director (PL)
49. Teyssou, Denis, Agence France-Presse, Head of AFP Medialab (FR)
50. Trommelen, Jeroen, Investico, Chief editor (NL)
51. Van Beek, Gijs, Textgain, Societal & Business innovator, Co-founder; Member of the Stakeholder Board (NL)
52. Van de Winkel, Goof, Graphic Hunters, Founder & owner (trainingsbureau for data visualisation and infographics) (NL)
53. Van Turnhout, Koen, Professor at Utrecht University of Applied sciences, Creative Business Programme (NL)

54. Vaudano, Maxime, Le Monde, Journalist, Les Décodeurs; Member of the Stakeholder Board (FR)
55. Wermuth, Mir, Journalism Fond, Investico, Media Perspective, freelance program manager in the media industries (NL)
56. Wilczyńska, Anna, Salam.Lab, Creator and owner (PL)
57. Witschage, Tamara, Amsterdam University of Applied Sciences, Professor of crossmedia (NL)
58. Wu, Shirley, Independent creator of data visualisations, freelance (USA)
59. Źaba, Natalia, Google News Initiative, Teaching Fellow (PL)
60. Zakrzewski, Patryk, Demagog.org.pl, Akademia Fact-Checking, coordinator of Fact-Checking (PL)

Literature

Adair, B. (2014, April 4). *Duke Study Finds Fact-Checking Growing Around the World*. Duke Reporter's Lab. https://reporterslab.org/duke-study-finds-fact-checking-growing-around-the-world/

AFP. (2022, September 19). *Artificial Intelligence versus Disinformation: AFP partner in the European project vera.ai.* Agence France-Presse press release. https://www.afp.com/en/agency/press-releases-newsletter/artificial-intelligence-versus-disinformation-afp-partner-european-project-veraai

AFP Australia. (2020, February 18). *This map shows flight paths worldwide -- it does not show the movement of Wuhan residents.* Agence France-Presse. https://factcheck.afp.com/map-shows-flight-paths-worldwide-it-does-not-show-movement-wuhan-residents

AFP, CORRECTIV, Pagella Politica/Facta, Maldita.es & Full Fact. (2020). *Infodemic Covid-19 in Europe: A Visual Analysis of Disinformation*. https://covidinfodemiceurope.com

Africa Check (n.d.). *What We Do: Training*. https://africacheck.org/what-we-do/training

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). *Evaluating the fake news problem at the scale of the information ecosystem.* Science Advances, 6(14). https://doi.org/10.1126/sciadv.aay3539

Apple Special Event. September 10, 2013. (2013, October 9). [Video]. YouTube. https://www.youtube.com/watch?v=yBX-KpMoxYk

Archer, H. (2021, March 15). *Online Media Literacy in Europe: Demand for Training is Going Unmet*. Ipsos. https://www.ipsos.com/ipsos-mori/en-uk/online-media-literacy-across-world-demand-training-going-unmet

Arockia, P., Varnekha, S., & Veneshia, K. (2017). *The 17 V's Of Big Data.* International Research Journal of Engineering and Technology (IRJET), 4(9), 330–331. https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf

Associated Press. (2018, May 30). *AP to automate video, audio transcription with Trint.* The Associated Press press release. https://www.ap.org/media-center/press-releases/2018/ap-to-automate-video-audio-transcription-with-trint/

Atkinson, C. (2019, April 25). *Fake news can cause 'irreversible damage' to companies — and sink their stock price*. NBC. https://www.nbcnews.com/business/business-news/fake-news-can-cause-irreversible-damage-companies-sink-their-stock-n995436

Baruch, J., Ferrer, M., Vaudano, M., & Michel, A. (2021, December 9). *OpenLux : the secrets of Luxembourg, a tax haven at the heart of Europe.* Le Monde. https://www.lemonde.fr/les-decodeurs/article/2021/02/08/openlux-the-secrets-of-luxembourg-a-tax-haven-at-the-heart-of-europe_6069140_4355770.html

Bauder, D. (2023, November. 29) *Sports Illustrated is the latest media company damaged by an AI experiment gone wrong.* The Associated Press https://apnews.com/article/journalists-ai-counterfeit-writers-479cc3869c0638df5bbb26d4b1e4f18f

Benkler, Y., Faris, R., and Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalisation in American Politics*. Oxford University Press.

Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., Faris, R., Kaiser, J., & Schmitt, C. (2020). *Mail-In Voter Fraud: Anatomy of a Disinformation Campaign*. Berkman Klein Center for Internet & Society at Harvard University. https://cyber.harvard.edu/publication/2020/Mail-in-Voter-Fraud-Disinformation-2020

Boykoff, M. T. & Boykof J. M. (2004). *Balance as bias: global warming and the US prestige press.* Global Environmental Change, 14, 125–136. https://www.eci.ox.ac.uk/publications/downloads/boykoff04-gec.pdf

Bradshaw, S., Bailey, H. & Howard, P. N. (2021). *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*. Computational Propaganda Project Working Paper Series. Oxford Internet Institute. University of Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/02/CyberTroop-Report20-Draft9.pdf

Brashier, N. M. & Marsh, E. J. (2020). *Judging Truth.* Annual Review of Psychology, 71(1), 499-515. https://doi.org/10.1146/annurev-psych-010419-050807

Breakstone, J., Smith, M., Connors, P., Ortega, T., Kerr, D., & Wineburg, S. (2021, February 23). *Lateral reading: College students learn to critically evaluate internet sources in an online course.* Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-56

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020, June 18). *The causes and consequences of COVID-19 misperceptions: Understanding the role of news*

and social media. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-028

Brown, J. (2019, July 9). *How social media could ruin your business*. BBC. https://www.bbc.co.uk/news/business-48871456

Buster, B. (2018). *Do Story*. Do Book Company.

Cairo, A. (2012). *The Functional Art: An Introduction to Information Graphics and Visualization* [E-book]. New Riders Publishing.

Cairo, A. (2019). *How Charts Lie: Getting Smarter about Visual Information* (Illustrated ed.) [E-book]. W. W. Norton & Company.

Cairo, A. (2016). *The Truthful Art: Data, Charts, and Maps for Communication (Voices That Matter)* (1st ed.) [E-book]. New Riders.

Cairo, A. & Rogers, S. (Hosts). (2021, April 27). *How do you judge data journalism?* (No. 1) [Audio podcast episode]. In The Data Journalism Podcast. https://open.spotify.com/episode/5fUvNR08Vxg5PDiYJoABdk

Cairo, A. & Rogers, S. (Hosts). (2021, May 14). *How to make data journalism for humans.* (No. 2) [Audio podcast episode]. In The Data Journalism Podcast. https://open.spotify.com/episode/05Zwlb4NHwVXWedQvC0rvf

Cairo, A. & Rogers, S. (Hosts). (2021 , June 8). *COVID data journalism special episode.* (No. 4) [Audio podcast episode]. In The Data Journalism Podcast. https://open.spotify.com/episode/0OFTcPOrnfG9f3zqJq3YeC

Cairo, A. & Rogers, S. (Hosts). (2021, Sept 28). *Eva Constanteras: doing data journalism in Afghanistan, Myanmar and across the Global South.* (No. 7) [Audio podcast episode]. In The Data Journalism Podcast. https://open.spotify.com/show/0vTWPzRFrkjfjy2lTFN4KC

Cairo, A. & Rogers, S. (Hosts). (2021, December 17). *2021 in data journalism: Scott Klein on how ProPublica does it, plus our favourite projects of the year.* (No. 9) [Audio podcast episode]. In The Data Journalism Podcast. https://open.spotify.com/episode/4AtFVNxEqNxGBMvgv8WBaa

Carrington, D. (2018, September 7). *BBC admits 'we get climate change coverage wrong too often.'* The Guardian. https://www.theguardian.com/environment/2018/sep/07/bbc-we-get-climate-change-coverage-wrong-too-often

Carrington, D. (2017, October 24). *BBC apologises over interview with climate denier Lord Lawson.* The Guardian. https://www.theguardian.com/environment/2017/oct/24/bbc-apologises-over-interview-climate-sceptic-lord-nigel-lawson

Chigwedere, P., Seage, G., Gruskin, S., & Lee, Tun-Hou. (2008). *Estimating the Lost Benefits of Antiretroviral Drug Use in South Africa*. JAIDS Journal of Acquired Immune Deficiency Syndromes, 49(4), 410-415. https://doi.org/10.1097/QAI.0b013e31818a6cd5

Chun, R. (2015, October 27). *6 lessons academic research tells us about making data visualizations*. Poynter. https://www.poynter.org/reporting-editing/2015/6-lessons-academic-research-tells-us-about-making-data-visualizations/

Clegg, N. (2019, September 24). *Facebook, Elections and Political Speech*. Facebook. https://about.fb.com/news/2019/09/elections-and-political-speech/

Collier, K. & Cui, J. (2024, February 4) *Why AI-generated audio is so hard to detect.* NBC https://www.nbcnews.com/tech/misinformation/ai-generated-audio-detect-tool-model-rcna136634

Cook, L. R. (2019, June 12). *How We Helped Our Reporters Learn to Love Spreadsheets*. Medium. https://open.nytimes.com/how-we-helped-our-reporters-learn-to-love-spreadsheets-adc43a93b919

Copeland, B.J. (2024, March 6). *artificial intelligence*. Britannica. https://www.britannica.com/technology/artificial-intelligence

Cunliffe-Jones, P, et al. (2021a). *Bad Law – Legal and Regulatory Responses to Misinformation in Eleven African Countries 2016–2020.* In Peter Cunliffe-Jones et al. (Eds.), *Misinformation Policy in Sub-Saharan Africa: From Laws and Regulations to Media Literacy.* (pp. 99–218). University of Westminster. https://doi.org/10.16997/book53.b

Cunliffe-Jones, P et al. (2021b). *The State of Media Literacy in Sub-Saharan African 2020 and a Theory of Misinformation Literacy*. In Peter Cunliffe-Jones et al. (Eds.), *Misinformation Policy in Sub-Saharan Africa: From Laws and Regulations to Media Literacy.* (pp. 5-96). University of Westminster. https://doi.org/10.16997/book53.a

Damgé, M., Michel, A., Vaudano, M., Baruch, J., & Ferrer, M. (2021, March 1). *OpenLux : enquête sur le Luxembourg, coffre-fort de l'Europe.* Le Monde.

https://www.lemonde.fr/les-decodeurs/visuel/2021/02/08/openlux-enquete-sur-le-luxembourg-coffre-fort-de-l-europe_6069132_4355770.html

Davies, P., (2023, December 15) *'Potentially disastrous' for innovation: Tech sector reacts to the EU AI Act saying it goes too far.* Euronews. https://www.euronews.com/next/2023/12/15/potentially-disastrous-for-innovation-tech-sector-says-eu-ai-act-goes-too-far

Demagog. (2021, February 27). *Nie, Jan Duda nie powiedział tego w wywiadzie dla RMF24*. Demagog. https://demagog.org.pl/fake_news/nie-jan-duda-nie-powiedzial-tego-w-wywiadzie-dla-rmf24/

Dixon, G. & Clarke, C. (2013). *The effect of falsely balanced reporting of the autism–vaccine controversy on vaccine safety perceptions and behavioral intentions.* Health Education Research, 28(2), 352–359. https://doi.org/10.1093/her/cys110

Dunlop, W. (2021, May 21). *US government database exploited by Covid-19 vaccine critics*. AFP Fact Check. https://factcheck.afp.com/us-government-database-exploited-covid-19-vaccine-critics

European Broadcasting Union (undated) *Promoting the interests of public service media.* https://www.ebu.ch/about

European Digital Media Observatory (n.d.) *EDMO Hubs*. https://edmo.eu/about-us/edmo-hubs/

European Digital Media Observatory (n.d.) *War in Ukraine.* https://edmo.eu/thematic-areas/war-in-ukraine/

European Commission, (2023, undefined) *Shaping Europe's digital future: AI Act.* https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#:~:text=The%20AI%20Act%20is%20the,play%20a%20leading%20role%20globally.&text=The%20AI%20Act%20aims%20to,regarding%20specific%20uses%20of%20AI

Fazio, L. K., Brashier, N. M., Payne, B. K., an&d Marsh, E. J. (2015). *Knowledge does not protect against illusory truth.* Journal of Experimental Psychology, 144(5), 993–1002. https://www.apa.org/pubs/journals/features/xge-0000098.pdf

Fitzgibbon, W. & Hudson, M. (2021, April 8). *Five years later, Panama Papers still having a big impact*. ICIJ. https://www.ICIJ.org/investigations/panama-papers/five-years-later-panama-papers-still-having-a-big-impact/

Full Fact. (2021). *The Full Fact Report 2021: Fighting a Pandemic Needs Good Information*. Full Fact. https://fullfact.org/media/uploads/full-fact-report-2021.pdf

Full Fact. (2020a, December 27). *Full Fact publishes new report on Facebook's Third-Party Fact-Checking programme*. Full Fact. https://fullfact.org/blog/2020/dec/full-fact-publishes-new-report-on-facebooks-third-party-fact-checking-programme/

Full Fact (2020b). *The Full Fact Report 2020: Fighting the Cause and Consequences of Bad Information*. Full Fact. https://fullfact.org/media/uploads/fullfactreport2020.pdf

Full Fact (2020c). *The Challenges of Online Fact Checking*. Full Fact. https://fullfact.org/media/uploads/coof-2020.pdf

Full Fact (n.d.). *Policy*. Full Fact. https://fullfact.org/about/policy/

Funke, D., (2019, June 18). *From Pants on Fire to Pinocchio: All the ways that fact-checkers rate claims*. Poynter Institute. https://www.poynter.org/fact-checking/2019/from-pants-on-fire-to-pinocchio-all-the-ways-that-fact-checkers-rate-claims/

Gabielkov, M., Ramachandran, A., Chaintreau, A. & Legout, A. (2016). *Social Clicks: What and Who Gets Read on Twitter?*. ACM SIGMETRICS / IFIP Performance 2016, Jun 2016, Antibes, Juan-les-Pins, France. https://hal.inria.fr/hal-01281190/document

Graves, L., & Mantzarlis, A. (2020). *Amid Political Spin and Online Misinformation, Fact Checking Adapts.* Political Quarterly 91(3), 585-591. https://doi.org/10.1111/1467-923X.12896

Grinberg, K. J., Friedland, L., Swire-Thompson, B. & Lazer, D. (2019, January 25). *Fake news on Twitter during the 2016 U.S. presidential election.* Science, 363(6425), 374-378. https://science.sciencemag.org/content/363/6425/374

Grynbaum, M. M., & Bromwich, J. E., (2021, March 26) *Fox News Faces Second Defamation Suit Over Election Coverage.* New York Times. https://www.nytimes.com/2021/03/26/business/media/fox-news-defamation-suit-dominion.html

Guess, R. (2024, January 10) *How AI Could Act as Boost for Investigative Journalism*. VOA New. https://www.voanews.com/a/how-ai-could-act-as-boost-for-investigative-journalism/7434364.html

Habgood-Coote, J. (2018, July 27). *The term 'fake news' is doing great harm*. The Conversation. https://theconversation.com/the-term-fake-news-is-doing-great-harm-100406

Hahn, J. (2021, June 22). *Blue bubbles helped "make the cause of climate change visible" say visualisers behind viral video*. Dezeen. https://www.dezeen.com/2021/06/22/carbon-real-world-visuals-new-york-emissions-interview/

Harrison Dupre, M. (2023, November 27) *Sports Illustrated Published Articles by Fake, AI-Generated Writers.* Futurism. https://futurism.com/sports-illustrated-ai-generated-writers

Holtz, Y. (2018). *The issue with pie chart*. Data to Viz. https://www.data-to-viz.com/caveat/pie.html

Hsu, S., (2023, October 14) *Analysis of cognitive warfare and information manipulation in the Israel-Hamas war 2023*, Taiwan AI Labs, https://ailabs.tw/uncategorized/analysis-of-cognitive-warfare-and-information-manipulation-in-the-israel-hamas-war-2023/

IBM - *What are large language models?* (n.d.). IBM. https://www.ibm.com/topics/large-language-models

IBM - *What is AI?* (n.d.). IBM. https://www.ibm.com/topics/artificial-intelligence

IBM - *What is deep learning?* (n.d.). IBM. https://www.ibm.com/topics/deep-learning

IBM - *What is generative AI?* (n.d.). IBM. https://research.ibm.com/blog/what-is-generative-AI

ICIJ. (n.d.). *Investigations Archives*. https://www.ICIJ.org/investigations/

J., D., & W. (2021). *CovidBaseAU | About*. CovidBaseAU. https://covidbaseau.com/about/

Johnson, B. (n.d.). *The Domesday Book*. Historic UK. https://www.historic-uk.com/HistoryUK/HistoryofEngland/Domesday-Book/

Jourova, V. (2023, September 26) *Press statement of Vice-President Jourova on the meeting with the Code of Practice on Disinformation Signatories. European Commission.* https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4645

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). *Motivated Numeracy and Enlightened Self-Government.* Behavioural Public Policy, 1(1), 54–86. Yale Law School, Public Law Working Paper No. 307. http://dx.doi.org/10.2139/ssrn.2319992

Kahan, D. (2011). *"What is Motivated Reasoning? How Does it Work?".* Discover Magazine. Blog Post reproduced at https://www.discovermagazine.com/the-sciences/what-is-motivated-reasoning-how-does-it-work-dan-kahan-answers

Kahn, J. P., & Damiano, M. (2021, September 22). *'They knew and they let it happen': Uncovering child abuse in the Catholic Church. Boston Globe.* https://www.bostonglobe.com/2021/09/22/magazine/they-knew-they-let-it-happen-uncovering-child-abuse-catholic-church/

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Books

Kelly, T. (Host). (2021, September). *Conversation with Antonio Baquero & Maxime Vaudano (ICIJ)* (No. 35) [Audio podcast episode]. In Conversations with Data. Data Journalism. https://soundcloud.com/datajournalism/episode-35-antonio-baquero-maxime-vaudano-occrp-le-monde

Kelly, T. (Host). (2021, October). *Conversation with Pierre Romera (ICIJ)* (No. 38) [Audio podcast episode]. In Conversations with Data. Data Journalism. https://soundcloud.com/datajournalism/episode-38-conversation-with-pierre-romera-ICIJ

Kelly, T. (Host). (2020, March). *Conversation with Craig Silverman* (No. 1) [Audio podcast episode]. In Conversations with Data. Data Journalism. https://soundcloud.com/datajournalism/datajournalismcom-podcast-with-craig-silverman

Kelly, T. (Host). (2020, March). *Conversation with Simon Rogers* (No. 3) [Audio podcast episode]. In Conversations with Data. Data Journalism. https://soundcloud.com/datajournalism/episode-3-simon-rogers-google

Kessler, G. (2020). Introduction: 16,000 Falsehoods. In Kessler, G., (Ed). *Donald Trump and His Assault on Truth: The President's Falsehoods, Misleading Claims and Flat-Out Lies.* Scribner.

Kim, Y., M. (2018, November 20). *Voter Suppression Has Gone Digital*. Brennan Center for Justice. https://www.brennancenter.org/our-work/analysis-opinion/voter-suppression-has-gone-digital

Krueger, V. (2017, June 8). *Here's what you should consider before using a fact-check rating system*. Poynter Institute. https://www.poynter.org/educators-students/2017/heres-what-you-should-consider-before-using-a-fact-check-rating-system/

Kunda, Z. (1990). *The case for motivated reasoning.* Psychological Bulletin, 108(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lacey, N. (2021, March 11). *COVID-19 vaccination intent has soared across the world.* Ipsos. https://www.ipsos.com/en/covid-19-vaccination-intent-has-soared-across-world

Lawrence F., Pegg, D. & Evans. R. (2019, October 10). *How vested interests tried to turn the world against climate science*. The Guardian. https://www.theguardian.com/environment/2019/oct/10/vested-interests-public-against-climate-science-fossil-fuel-lobby

Leffer, L. (2024, March 4) *Everything to Know About OpenAI's New Text-to-Video Generator, Sora.* Scientific American. https://www.scientificamerican.com/article/sora-openai-text-video-generator/

Le Parisien (2024, March 5) *Affaire Depardieu : l'acteur visé par une nouvelle enquête pour agression sexuelle, après la plainte d'une décoratrice.* Le Parisien. https://www.leparisien.fr/faits-divers/affaire-depardieu-lacteur-vise-par-une-nouvelle-enquete-pour-agression-sexuelle-apres-la-plainte-dune-decoratrice-05-03-2024-IXHZE4SETFFOXGIHNGDNDVHA7E.php

Levitin, D. (2018). *A Field Guide to Lies and Statistics: A Neuroscientist on How to Make Sense of a Complex World*. Penguin.

Lichtenstein, S., Slovic, P., Fischhoff, B., & Layman, M. (1978.) *Judged Frequency of Lethal Events.* Journal of Experimental Psychology Human Learning and Memory, 4(6), 551-578. https://doi.org/:10.1037/0278-7393.4.6.551

Lipton, E. (2016, December 5). *Man Motivated by 'Pizzagate' Conspiracy Theory Arrested in Washington Gunfire.* New York Times. https://www.nytimes.com/2016/12/05/us/pizzagate-comet-ping-pong-edgar-maddison-welch.html

Loguercio, L., and Canepa, C. (2021). *Fact-Checking Engagement Project*. Pagella Politica. https://www.engagingwithfacts.org/wp-content/uploads/2021/04/FCEP_Handbook.pdf

Loomba, S., de Figueiredo, A., Piatek, S.J., de Graaf, K., & Larson, H. J., (2021). *Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA.* Nat Hum Behav, 5, 337–348. https://doi.org/10.1038/s41562-021-01056-1

Makortoff, K., Brignall, M., & Waterson, J. (2019, May 13). *Metro Bank shares plunge as it attacks 'false rumours.'* The Guardian. https://www.theguardian.com/business/2019/may/13/metro-bank-shares-rumours-safety-deposit-boxes

Maldita.es. (2018, August 28). *Maldita.es: journalism to not be fooled*. Maldita. https://maldita.es/maldita-es-journalism-to-not-be-fooled

Mantas, H. (2021, May 6). *Factually: Fact-checkers advocate for an end to Facebook's ban of fact-checking political figures*. Poynter Institute. https://www.poynter.org/fact-checking/2021/factually-fact-checkers-advocate-for-an-end-to-facebooks-ban-of-fact-checking-political-figures/

Mantzarlis, A. (2019, December 19). *How we highlight fact checks in Search and Google News*. Google News Initiative. https://blog.google/outreach-initiatives/google-news-initiative/how-we-highlight-fact-checks-search-and-google-news/

Mantzarlis, A. (2018). MODULE 5: Fact-checking 101. In Ireton, C., and Posetti, J., (Eds), *Journalism, 'Fake News' & Disinformation: Handbook for Journalism Education and Training*. United Nations Educational, Scientific and Cultural Organization. (pp. 85-100). https://unesdoc.unesco.org/ark:/48223/pf0000265552

Mark, S., & Luther, J. (2020, June 22). *Annual census finds nearly 300 fact-checking projects around the world*. Duke Reporter's Lab. Duke University. https://reporterslab.org/latest-news/

Martel, C,. Pennycook G., & Rand, D. (2020). *Reliance on emotion promotes belief in fake news.* Cognitive Research, 5(47). https://doi.org/10.1186/s41235-020-00252-3

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

McCready, R. (2021, November 17). *5 Ways Writers Use Misleading Graphs To Manipulate You* [INFOGRAPHIC]. Venngage. https://venngage.com/blog/misleading-graphs/

McKinsey & Company. (2023, January 19) *What is generative AI?* McKinsey & Company https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

Mearian, L. (2024, February 7.) *What are LLMs, and how are they used in generative AI?* Computer World https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html

Nguyen, K., Baris-Schlicht, I., Altiok, D., Mortada, S., Waleed, K., Taouk, M. (2023, September 18) *BiasBlocker: We asked a language model to identify racism and it tried to erase baby Hitler.* JournalismAI. https://www.journalismai.info/blog/we-asked-a-language-model-to-identify-racism-and-it-tried-to-erase-baby-hitler

Nur, F. (2019, October 2). *The rumour that led to medical researchers in Ethiopia being killed by a mob*. BBC. https://www.bbc.co.uk/programmes/p07pvjxx

Nyhan, B. (2020) *Facts and Myths About Misperceptions.* The Journal of Economic Perspectives, 34(4), 220- 236. https://www.jstor.org/stable/10.2307/26923548

Nyhan, B., Porter, E., Reifler, J., & Wood. T. J. (2019). *Taking Fact-Checks Literally but Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability.* Political Behavior, 42, 939-960. https://doi.org/10.1007/s11109-019-09528-x

Nyhan, B, & Reifler, J. (2015). *The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators.* American Journal of Political Science, 59(3), 628-640. https://www.jstor.org/stable/24583087

Nyhan, B, & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior 32(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

O'Brien, M. (2023, July 13) *ChatGPT-maker OpenAI signs deal with AP to license news stories.* Associated Press https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a

Osmundsen, M., Bor, A., Vahlstrup, P., Bechmann, A., & Petersen, M. (2021). *Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter*. American Political Science Review, *115*(3), 999-1015. https://doi.org/10.1017/S0003055421000290

Outside in America team. (2017, December 20). *Bussed out: how America moves thousands of homeless people around the country.* The Guardian. https://www.theguardian.com/us-news/ng-interactive/2017/dec/20/bussed-out-america-moves-homeless-people-country-study

Pennycook, G. & Rand, D. G. (2021). *The Psychology of Fake News.* Trends in Cognitive Sciences 25(5). https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand D. G. (2021, April 22). *Shifting attention to accuracy can reduce misinformation online.* Nature, 592, 590-616. https://doi.org/10.1038s41586-021-03344-2

Pennycook, G. & Rand, D. G. (2019). *Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning*. Cognition, 188, 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Pivot, B. (2019, December 19) *1990 : Gabriel Matzneff face à Denise Bombardier dans "Apostrophes"*. INA Archive. https://www.youtube.com/watch?v=H0LQiv7x4xs

Ponsford, D. (2023, November 16) *Newsquest CEO Henry Faure Walker on bucking the trend of regional press decline.* Press Gazette. https://pressgazette.co.uk/publishers/regional-newspapers/newsquest-ceo-henry-faure-walker-on-bucking-the-trend-of-regional-press-decline/

Porter, E., Velez, Y., & Wood, T. J. (2021). *Factual Corrections Eliminate False Beliefs About COVID-19 Vaccines*. OSF. https://doi.org/10.17605/OSF.IO/P47BT

Poujol, V., (2022, November 24) *Coup dur pour la transparence financière*, Reporter. https://www.reporter.lu/luxembourg-cour-de-justice-ue-coup-dur-pour-la-transparence-financiere

Poynter Institute. (2020). *State of Fact-Checking 2020.* Poynter Institute. https://www.poynter.org/wp-content/uploads/2020/06/IFCN_2020_state-of-fact-checking_ok.pdf

Rahman, G. (2020a, April 9), *Here's where those 5G and coronavirus conspiracy theories came from*. Full Fact. https://fullfact.org/online/5g-and-coronavirus-conspiracy-theories-came/

Rahman, G. (2020b, March 5). *Viral post about someone's uncle's coronavirus advice is not all it's cracked up to be*. Full Fact. https://fullfact.org/online/coronavirus-claims-symptoms-viral/

Ravitz, J. (2018, September 5). *Gwyneth Paltrow's Goop brand hit with penalties for 'unsubstantiated claims'*. CNN. https://edition.cnn.com/2018/09/05/health/goop-fine-california-gwyneth-paltrow/index.html

Rogers, S. (2014, May 29). *Introduction to data journalism*. Simon Rogers. https://simonrogers.net/2014/05/25/introduction-to-data-journalism/

Roozenbeek, J., and van der Linden, S. (2019). *Fake news game confers psychological resistance against online misinformation.* Palgrave Commun, 5(65). https://doi.org/10.1057/s41599-019-0279-9

Roper, D. (2023, July 7) *Michael Miller on how NewsCorp Australia has transformed its journalism and business.* World Association of News Publishers https://wan-ifra.org/2023/07/michael-miller-on-how-newscorp-australia-has-taken-a-stand-and-transformed-its-journalism-and-business/

Rosen, G. (2021, March 22). *How We're Tackling Misinformation Across Our Apps.* Reproduced on Facebook. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/

Rothman, M., & Miller, D. (2016, February 29). *The Real "Spotlight": Meet Team That Inspired the Oscar-Winning Film*. ABC News.

https://abcnews.go.com/Entertainment/real-spotlight-meet-team-inspired-oscar-nominated-film/story?id=37139332

Safi, M. (2018, July 3). *'WhatsApp murders': India struggles to combat crimes linked to messaging service.* The Guardian. https://www.theguardian.com/world/2018/jul/03/whatsapp-murders-india-struggles-to-combat-crimes-linked-to-messaging-service

Scheufele, D. A., & Krause, N. M. (2019). *Science audiences, misinformation, and fake news* (Vol. 116, Issue 16). https://doi.org/10.1073/pnas.1805871115

Senate Intelligence Committee. *Report of the Select Committee on Intelligence on Active Measures Campaigns and Interference in the 2016 U.S. Election*, 116-XX. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf

Serokell, S. (2021, December 26). *What Is Big Data?* Medium. https://ai.plainenglish.io/what-is-big-data-646d5e5bedc3

Silverman, C. (2016, November 16). *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook*. BuzzFeed. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Sippitt, A., & Moy, W. (2020). *Fact Checking is About What We Change Not Just Who We Reach.* The Political Quarterly, 91(3), 592–595. https://doi.org/10.1111/1467-923x.12898

Sippitt, A. (2019, March 20). *The Backfire Effect: Does it Exist? And Does It Matter To Factcheckers?* Full Fact. https://fullfact.org/blog/2019/mar/does-backfire-effect-exist/

Sivadas, L. (2024, March 11) *Welcoming the 2024 cohort of JournalismAI Fellows.* JournalismAI. https://www.journalismai.info/blog/welcoming-the-2024-cohort-of-journalismai-fellows

Stecula, D. A., Kuru, O., & Jamieson, K. H. (2020). *How trust in experts and media use affect acceptance of common anti-vaccination claims.* Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-007

Storr, W. (2020). *The Science of Storytelling*. Macmillan Publishers.

Swift, J. (1710, November 9). *The Examiner*, 14.

Swire, B., Berinsky, A. J., Lewandowsky, S. & Ecker U. K. H. (2017). *Processing Political Misinformation: Comprehending the Trump Phenomenon.* Royal Society Open Science, 4(3). http://dx.doi.org/10.1098/rsos.160802

Tapestry 2015 Short Stories - Ben Jones: "Seven Data Story Types." (2015, March 13). [Video]. YouTube. https://www.youtube.com/watch?v=sEZj-eUfbNo&feature=youtu.be

TEMS. *Trusted European Media data Space*. TEMS https://tems-dataspace.eu/about/

Thompson, T. (2023, November 10) *No evidence clip of Sadiq Khan supposedly calling for 'Remembrance weekend' to be postponed is genuine.* Full Fact. https://fullfact.org/news/khan-audio-palestinian-remembrance/

Trafton, A. (2014, January 16). *In the blink of an eye*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2014/in-the-blink-of-an-eye-0116

Tufte, R. E. (2021). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press.

Yanofsky, D. (2020, June 24). *The chart Tim Cook doesn't want you to see*. Quartz. https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

VAERS. *Data*. (n.d.). VAERS. https://vaers.hhs.gov/data.html

Van Duyn, E., & Collier, J., (2019) *"Priming and Fake News: The Effects of Elite Discourse on Evaluations of News Media."* Mass Communication and Society 22(1), 29-48. https://doi.org/10.1080/15205436.2018.1511807

Vandewalker, I. (2020, September 2). *Digital Disinformation and Vote Suppression*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/digital-disinformation-and-vote-suppression

Vosoughi, S., Deb, R., & Aral, S., (2018). *The Spread of True and False News Online.* Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Walter, N., & Tukachinsky, R. (2019). *A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?* Communication Research, 47(2),155-177. https://doi.org/10.1177/0093650219854600

Wardle, C., & Derakhshan, H. (2018). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe. https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77

Wardle, C. (2017). *Fake News. It's Complicated*. First Draft. February 16, 2017. https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

Wilkinson, K. (2018, September 2). *ANALYSIS: Crime rates worse than South African police calculated*. Africa Check. https://africacheck.org/fact-checks/blog/analysis-crime-rates-worse-south-african-police-calculated

Yang, W. (2024, January 5) *Q&A: Taiwan AI Labs Founder Warns of China's Generative AI Influencing Election*. VOA News https://www.voanews.com/a/q-a-taiwan-ai-labs-founder-warns-of-china-s-generative-ai-influencing-election-/7428717.html

Zeng, E., Kohno, T., & Roesner, F. (2020). *Bad News: Clickbait and Deceptive Ads on News and Misinformation Websites*. Workshop on Technology and Consumer Protection. https://homes.cs.washington.edu/~yoshi/papers/ConPro_Ads.pdf

Zimmermann, F., & Kohring, M. (2020). *Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election.* Political Communication, 37(2), 215–237. https://doi.org/10.1080/10584609.2019.1686095

Figure Credits

Figure 1. *Have you heard of Ishango?* (n.d.). Royal Belgian Institute of Natural Sciences. https://www.naturalsciences.be/sites/default/files/Discover%20Ishango.pdf

Figure 2. *Diagram of the causes of mortality in the Army in the East - Digital Collections - National Library of Medicine*. (n.d.). National Library of Medicine. https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101598842-img

Figure 3. *Boston Globe*. (2002, January 6). *Sexual abuse in the Catholic Church* [Screenshot]. Boston Globe. https://www.bostonglobe.com/metro/investigations/spotlight/?p1=Article_Inline_Text_Link

Figure 4. ICIJ. (2017, January 31). *The Panama Papers: Exposing the Rogue Offshore Finance Industry* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/panama-papers/

Figure 5. ICIJ. (2017, November 5). *Paradise Papers: Secrets of the Global Elite* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/paradise-papers/

Figure 6. ICIJ. (2018, November 25). *Implant Files* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/implant-files/

Figure 7. ICIJ. (2021a, October 3). *Pandora Papers* [Screenshot]. ICIJ. https://www.ICIJ.org/investigations/pandora-papers/

Figure 8. Outside in America team, Bremer, N., & Wu, S. (2017, December 20). *Homeless bus relocation journeys to destinations in the mainland US* [Screenshot]. The Guardian. https://www.theguardian.com/us-news/ng-interactive/2017/dec/20/bussed-out-america-moves-homeless-people-country-study

Figure 9. Real World Visuals. (2012, October 19). *New York City's greenhouse gas emissions as one-ton spheres of carbon dioxide gas* [Video]. YouTube. https://www.youtube.com/watch?v=DtqSIplGXOA

Figure 10. Holtz. Y. (2018). *Comparison of pie charts versus bar plots* [Graph]. Data to Viz. https://www.data-to-viz.com/caveat/pie.html

Figure 11. *Apple Special Event. September 10, 2013.* (2013, October 9). [Video]. YouTube. https://www.youtube.com/watch?v=yBX-KpMoxYk

Figure 12. Yanofsky, D. (2020, June 24). *The chart Tim Cook doesn't want you to see*. [Screenshot] Quartz. https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

Figure 13. Cairo, A. (2020, January 8). *All graphics from "How Charts Lie" freely available in two color schemes*. The Functional Art. http://www.thefunctionalart.com/2020/01/all-graphics-from-how-charts-lie-freely.html

Figure 14. Cairo, A. (2020, January 8). *All graphics from "How Charts Lie" freely available in two color schemes*. The Functional Art. http://www.thefunctionalart.com/2020/01/all-graphics-from-how-charts-lie-freely.html

Figure 15. Stock images of red traffic signs

Figure 16. Map showing climate action announced by 37 countries and the European Union following the COP26 climate summit, according to Climate Action Tracker. Source: AFP/Valentina Breschi, Kun Tian
https://www.afpforum.com/ ID number 9RM2LF

Figure 17. Britain's Catherine, Duchess of Cambridge holds Britain's Prince Louis of Cambridge on their arrival for his christening service at the Chapel Royal, St James's Palace, London on July 9, 2018. Source: POOL/AFP/Dominic Lipinski
https://www.afpforum.com/ ID number 17E44Q

Figure 18. Dressed in traditional Korean mourning white, the sisters of Park Mi-Jin, one of scores of young salesgirls killed in the collapse of the Sampoong Department Store, help their grieving mother (C) to the funeral on 3 July near Seoul's Kangnam Hospital. More than 200 are still missing beneath the rubble of the store. Source: AFP/KIM JAE-HWAN
https://www.afpforum.com/ ID number APW2002052949691

Figure 19. Chart showing annual fossil carbon dioxide emissions and 2021 projections, according to the Global Carbon Project 2021. Source: AFP/Valentina Breschi, Gal Roma
https://www.afpforum.com/ ID number 9RG4XZ and 9RH4JL

Figure 20. Dunlop, W. (2021, May 21). *US government database exploited by Covid-19 vaccine critics.* AFP Fact Check.
https://factcheck.afp.com/us-government-database-exploited-covid-19-vaccine-critics

Figure 21. Screenshot of a since-removed tweet by the WorldPop Project, as recorded by AFP Factcheck. AFP Australia (2020, February 18). *This map shows flight paths worldwide -- it does not show the movement of Wuhan residents.* AFP Factcheck
https://factcheck.afp.com/map-shows-flight-paths-worldwide-it-does-not-show-movement-wuhan-residents

Archived screenshot of the tweet can be found here:
https://web.archive.org/web/20200211004916/https:/twitter.com/WorldPopProject/status/1225132600420917254

13. Appendix

Appendix I: Suggested Resources

This is a suggested list of useful resources but it is by no means an exhaustive list.

Literature

Beckett, C., Yaseen, M. (2023, September 20) *Generating Change. A global survey of what news organisations are doing with AI.* Polis. [downloadable pdf] https://www.journalismai.info/research/2023-generating-change

Bradshaw, P. (2013). *Scraping for Journalists.* E-book

Cairo, A. (2020). *How Charts Lie: Getting Smarter About Visual Information*. W. W. Norton & Company. http://www.thefunctionalart.com/p/reviews.html

Cairo, A. (2012). *The Functional Art, The: An introduction to information graphics and visualization* (Voices That Matter). (1st ed.). New Riders. http://www.thefunctionalart.com/p/about-book.html

Cairo, A. (2016). *The Truthful Art: Data, Charts, and Maps for Communication* (1st ed.). New Riders. http://www.thefunctionalart.com/p/the-truthful-art-book.html

Dykes, B. (2019) *Effective Data Storytelling*. Wiley. https://www.effectivedatastorytelling.com/

Gray, J., Bounegru, L., *The Data Journalism Handbook 2* (2020) European Journalism Centre [downloadable pdf] https://datajournalism.com/read/handbook/two

European Journalism Centre (2023) *The State of Data Journalism 2022*. European Journalism Centre. [downloadable pdf] https://datajournalism.com/survey/2022/

Keng, K., Ser, K. (n.d.), *Best Practices for Data Journalism* (n.d.) [downloadable pdf] https://www.kbridge.org/wp-content/uploads/2018/04/Guide-3-Best-Practices-for-Data-Journalism-by-Kuang-Keng.pdf

Newman, N., Fletcher, R., Eddy, K., Robertson, C.T., Kleis Nielsen, R. (2023, June 14) *Reuters Institute Digital News Report 2023*. Reuters Institute, University of Oxford. [downloadable pdf] https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/dnr-executive-summary

Nussbaumer Knaflic, C. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals* (1st ed.). Wiley. https://www.storytellingwithdata.com/books

Nussbaumer Knaflic, C. (2019). *Storytelling with Data: Let's Practice!* (1st ed.). Wiley. https://www.storytellingwithdata.com/books

Rogers, S. (2013). *Facts are Sacred*. Guardian Faber Publishing

Silverman, C. (n.d.) *Verification Handbook: For Disinformation and Media Manipulation*. European Journalism Centre [downloadable pdf] https://datajournalism.com/read/handbook/verification-3

Tufte, R. E. (2021). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press https://www.edwardtufte.com/tufte/books_vdqi

Vo, L, T (2019) *Mining Social Media*. Penguin Random House https://lamivo.com/work.html#books

Videos

Explore Explain from Andy Kirk. https://www.youtube.com/channel/UCIPsLvCpZYwvSurkb1DLLZg

JournalismAI Festival. https://www.youtube.com/@PolisLSE

Storytelling with Data from Cole Nussbaumer Knaflic. https://www.youtube.com/channel/UCjhGlILWNloXJdR2NTCBMlA

Podcasts

Data Journalism
Conversations with Data from Tara Kelly. https://datajournalism.com/listen

Data Stories from Enrico Bertini and Moritz Stefaner. https://datastori.es/

Data Viz Today from Alli Torban.
https://dataviztoday.com/shownotes

Explore Explain from Andy Kirk.
https://pod.co/exploreexplain

Loud Numbers from Miriam Quick and Duncan Geere.
https://www.loudnumbers.net/

Storytelling with Data from Cole Nussbaumer Knaflic.
https://www.storytellingwithdata.com/podcast

The Data Journalism Podcast from Alberto Cairo and Simon Rogers.
https://anchor.fm/ddjpodcast

Digital Verification
Full Fact podcast (UK)
https://open.spotify.com/show/3KH3Aaf3Rq72tSNpaGRVMj

Infox, France Culture.
https://www.franceculture.fr/theme/fake-news

Observador (Portugal) factcheck podcast
https://open.spotify.com/show/1DCIgrLCPyiugkHCH3HLyJ

Maldita podcast “Maldita la hora” (Spain)
https://open.spotify.com/show/5zWm9bW6SFxx11a1AOjhtJ?si=oUftcc-2TPStwoJEM4V6vg&nd=1

Pagella Politica podcast (Italy)
https://open.spotify.com/show/0uOO1kCUWQifJSPMgTLSPz

Best and Worst Case Examples

Inspiring data journalism
Sigma Awards
https://sigmaawards.org/

Examples of Poor Practice in Data Visualisation
Heap: How to lie with data visualization
https://heap.io/blog/how-to-lie-with-data-visualization

Medium: Bad Data Visualization in the Time of Covid-19
https://medium.com/nightingale/bad-data-visualization-in-the-time-of-covid-19-5a9f8198ce3e

Reddit: Data is Ugly

https://www.reddit.com/r/dataisugly/

Tumblr: Bad Visualizations
https://badvisualisations.tumblr.com/

Venngage: 5 Ways writers use misleading graphs to manipulate you
https://venngage.com/blog/misleading-graphs/

Venngage: The Worst Infographics of 2020
https://venngage.com/blog/bad-infographics/

Wikipedia: Misleading graph
https://en.wikipedia.org/wiki/Misleading_graph

Useful Links

Data
Datajournalism.com
https://datajournalism.com/

European Data Journalism Network
https://www.europeandatajournalism.eu/

European Journalism Centre
https://ejc.net/training

Journalism AI
https://www.journalismai.info/

MediaNumeric Academy
https://www.medianumericacademy.eu/

Sigma Awards
https://sigmaawards.org/

Digital Verification
Claim Review Project: a way of labelling fact-check articles
https://www.claimreviewproject.com/

The Content Authenticity Initiative (CAI) launched by Adobe
https://contentauthenticity.org/

Duke Reporters Lab
https://reporterslab.org/fact-checking/
https://factcheck.afp.com/how-find-source-video-or-how-do-reverse-video-search

EDMO
https://edmo.eu/

European Fact-Checking Standards Network (EFCSN)
https://efcsn.com/

International Fact-Checking Network (IFCN)
https://www.poynter.org/ifcn/

Journalism AI
https://www.journalismai.info/

MediaNumeric Academy
https://www.medianumericacademy.eu/

The Partnership on AI
https://partnershiponai.org/

Politwoops: tracks deleted tweets by public officials
https://projects.propublica.org/politwoops/

The Virality Project: fighting online COVID-19 vaccine disinformation
https://www.viralityproject.org/

Tools

There are many tools that are free to use and tutorials are widely and freely available online in order to learn how to use them. Need tools are released onto the market regularly so this list is not exhaustive.

Data Visualisations:

D3
https://d3js.org/

DataWrapper
https://www.datawrapper.de/

Dataiku
https://www.dataiku.com/

Figma
https://www.figma.com/

Flourish

https://flourish.studio/

Localfocus: data visualization
https://www.localfocus.nl/en/

R
https://www.r-project.org/

Tableau
https://www.tableau.com/

Programming:
Python: considered one of the easiest programming languages for a beginner to learn.
https://www.python.org/

R
https://www.r-project.org/

Text Search Inside Files:
Datashare: free open-source desktop application which allows journalists to simultaneously search pdfs, images, texts, slides or any other files. It can also automatically detect and filter by people, organisations and locations. The ICIJ also secures documents from third-party interference.
https://datashare.icij.org/

Pdftotext: open source toolkit created by Xpdf which can also be used to extract text from pdf files.
https://www.xpdfreader.com/pdftotext-man.html
https://www.xpdfreader.com/

Pinpoint
https://journaliststudio.google.com/pinpoint/about/

Digital verification tools:
Africa Check Info Finder: sources reliable info on African topics
https://africacheck.org/infofinder

CrowdTangle: analyses facebook posts
https://www.crowdtangle.com/

Google: reverse image search
https://www.google.com/imghp?hl=en

InVid-WeVerify: Plug-in toolbox for analysing photos and videos

https://www.invid-project.eu/
https://weverify.eu/
https://www.afp.com/en/agency/medialab/invid
https://www.veraai.eu/home

Tineye: reverse image search
https://tineye.com/

Audio transcription:
CQ Transcripts
https://info.cq.com/legislative-news/cq-transcripts-testimony/

Chequeado's Chequeabot: transcription tool
https://chequeabot.chequeado.com/transcriptor/

Rev transcriptions
https://www.rev.com/

Trint
https://trint.com/

Whisper
https://openai.com/research/whisper

AI-powered chatbots:
ChatGPT: AI-powered chatbot
https://chat.openai.com/auth/login

Gemini: AI-powered chatbot
https://gemini.google.com/

AI-powered text-to-Image generation:
Adobe Firefly
https://www.adobe.com/products/firefly.html

DALL-E: AI-powered text-to-image generator
https://openai.com/dall-e-2

Ideogram
https://ideogram.ai/login

Leonardo

https://leonardo.ai/

Midjourney
https://www.midjourney.com/home

Runway
https://runwayml.com/

AI-powered text-to-video generation:
Pika
https://pika.art/home

Sora
https://openai.com/sora

Other:
BuzzSumo: monitoring online trends
https://buzzsumo.com/

GIMP: The open source image editor
https://www.gimp.org/

Newswhip: social media analytics
https://www.newswhip.com/

X (formerly Twitter) Accounts

Data
Alberto Cairo @albertocairo
Andy Kirk @visualisingdata
Christina Elmer @ChElm
Cole Nussbaumer Knaflic@storywithdata
David Cabo @dcabo
Data Driven Journalism @ddjournalism
David McCandless@mccandelish
Eva Constantaras @EvaConstantaras
Gurman Bhatia @GurmanBhatia
Giorgia Lupi @giorgialupi
Gregor Aisch @driven_by_data
ICIJ@ICIJorg

Josh Holder @Josh_H
Les Décodeurs@decodeurs
Martin Stabe @martinstabe
Maxime Vaudano@mvaudano
Mona Chalabi @MonaChalabi
Nadieh Bremer @NadiehBremer
Óscar Marín @oscarmarinmiro
Paul Bradshaw @paulbradshaw
Robert Kosara @eagereyes
Sam Joiner @samjoiner
Shirley Wu @sxywu
sigmaawards@sigmaawards
Simon Rogers @smfrogers
Zete Hausfather @hausfath

Debunking Misinformation
AFP Fact Check @AFPFactCheck
Baybars Örsek @baybarsorsek
Angie Holan @AngieHolan
Full Fact @FullFact
Glenn Kessler @GlennKesslerWP
IFCN @factchecknet
Peter Cunliffe-Jones @PCunliffeJones
PolitiFact @PolitiFact
Sophie Nicholson @sohnic
Will Moy @puzzlesthewill

Appendix II: MediaNumeric Consortium partners

Netherlands Institute for Sound and Vision (NISV), The Netherlands
www.beeldengeluid.nl/en

Sound and Vision is the leading institute for media in the Netherlands and one of the largest and trendsetting audiovisual archives in the world. It is an inspiring, creative and welcoming meeting place for professionals and others interested in the industry; online, in its physical museum and on location, for instance at festivals. NISV preserves and provides access to different types of media, including radio and television programmes, video games, written print media, political cartoons, GIFs, websites and historical objects. It is one of the leading authorities when it comes to providing insight into the Dutch media landscape and interpreting current developments from the perspective of media history. NISV does this with the aim of showing how media impacts everyday life and does so in conjunction with many partners, including creative media makers, key experts, business stakeholders and relevant influencers. In this way, NISV helps to build a more media literate world.

Inholland University of Applied Sciences (INH), The Netherlands
https://www.inholland.nl/inhollandcom

Inholland University of Applied Sciences is an institution of higher education in the western part of the Netherlands. Its organisational structure is a result of a collaboration of four formally independent institutions. Across nine campuses, Inholland offers 80 bachelor programmes in all fields of study: from journalism to economics, from technology to law. In addition, INH is home to seven master programmes and is in the top five universities of applied sciences, with its media bachelor's programme being the best in the Netherlands. Students at Inholland combine theoretical knowledge with practical learning through work on real-life projects. The Creative Business department offers six bachelor programmes in the creative industries (over 6,000 students) and is the home of the Creative Business research group. The group has a long tradition of audience research in media and cultural studies and practice-based research in innovations in professional cultures in the creative industries, media literacy in a digital age, and inclusive communication.

University of Social Sciences and Humanities (SWPS), Poland
www.english.swps.pl

SWPS University of Social Sciences and Humanities (SWPS University) is a leading private higher education institution in Poland specialising in social sciences, including media studies. The University was established in 1996 and now is a community of 400 permanent faculty of researchers and experienced academics who teach over 14,000 students enrolled in 35 undergraduate, graduate and doctoral programmes. The broad education offer includes 15

programmes taught entirely in English to over 1,200 international students from more than 60 countries. SWPS University is one of the most active research universities in Poland also focusing its research interests on the topic of media ecosystem changes and its impact on society.

Centrum Cyfrowe (CC), Poland
https://centrumcyfrowe.pl/en/homepage/

Centrum Cyfrowe works to make the world more inclusive, more cooperative and more open by changing the way people learn, participate in culture, use the internet and exercise their rights as internet users. They support users of digital technologies in improving their skills and competences related to openness and cooperation and cooperate with institutions to make sure that they work in an open manner in order to carry out their social mission. They also work towards adjusting regulations and using legal tools to support the needs and rights of users and diagnose social and cultural changes taking place in our society with the influence of digital technologies.

Agence France-Presse (AFP), France
https://www.afp.com/en
https://factcheck.afp.com/

Agence France-Presse (AFP) is a global news agency, delivering verified news worldwide and fast, accurate, in-depth coverage of the events shaping our world from wars, conflicts to politics, sports, entertainment and the latest breakthroughs in health, science or technology. With 2,300 staff of 80 nationalities, spread across 165 countries, AFP covers the world 24 hours a day, producing about 3,000 articles, 3,000 photos, 100 graphics and 250 videos per day in six languages. Since 2017, AFP has built up the first global digital investigations network with more than 130 journalists, currently covering 85 countries in 24 languages.

Institut national de l'audiovisuel (INA), France
https://institut.ina.fr/en

INA is a French public institution with industrial and commercial purposes, entrusted with the responsibility of the archiving, preservation and access of French audiovisual heritage, and the development and transmission of knowledge in this field. INA is in charge of the Audiovisual Legal Deposit: the 24/7 collection of 168 radio and TV channel programmes, raising the volume of its collections to 18 million hours. It also collects the content of websites and Twitter accounts dealing with media. To give access to its collections, INA has developed offers dedicated to specific audiences: a videoclips website and social media contents for the general public, an online service dedicated to professionals, consultation centres for academic purposes, and multimedia tools ready-to-use by teachers and students for educational uses.

To transmit its knowledge, INA is a professional training and higher education training centre. It offers a professional training catalogue as well as custom-made training programmes in archive,

media and digital sectors. With INAsup, it provides 14 training courses, awarding graduate and postgraduate diplomas, in all audiovisual and media fields.

Storytek (ST), Estonia
https://storytek.eu/

Storytek is the first personalised Hollywood-grade creative/mediatech/and storytelling accelerator in Northern Europe. Founded in 2017 by private investors and the Tallinn Black Nights Film Festival, Storytek brings together deep audiovisual sector knowledge, technology and funding with a selection of storytellers and media and technology entrepreneurs. Through a 10-week intensive masterclass programme culminating with mentors across the world's top media and technology companies, Storytek helps storytellers and entrepreneurs develop their projects into pilots by boosting skills from creative and storytelling to business models, financing, and sales and distribution strategies ready to be pitched at Storytek's investor and demo day.

Besides project development Storytek also advises several media and tech companies from telcos to broadcast integrators on new content and investment projects, organises training and innovation-boosting educational events, works on several European and local R&D projects (including Horizon 2020) as well as provides consultation and advisory from feature and documentary projects to online, interactive and feature content to startups. In the past two years Storytek has boosted 24 projects from 9 countries, and pushed them to international focus from Berlinale EFM to international investors.

Appendix III: Partner Input WP2 - Instructions

Needs Analysis & State of the Art Analysis

Deadline: February 5, 2021

—

The aim:
The updated NEEDS ANALYSIS and updated STATE OF THE ART ANALYSIS focus on both data-driven journalism and data-driven work in the creative and media industries.

For the NEEDS ANALYSIS:
It means collecting data about required transversal skills, adaptive digital competencies, and critical thinking skills that ensure the employability of future professionals. The needs assessment should be conducted from both perspectives -- stakeholders of the media and creative industry, and from the side of the relevant Higher Education fields.

For the STATE OF THE ART ANALYSIS:
It includes the collection of best practices, challenges (including lessons learnt from failed projects), trends as collected via expert interviews and desk research. It also includes the inventory of large data sets used in Higher Education and/or suitable for HE and by media and creative industries professionals more broadly. It maps innovative practices used to link journalism and professional communication with work on data, supported also by technology (AI) and identifies current approaches (trends and gaps) in teaching methods. It includes a survey and mapping of the state of the art detailed and divided by partner country.
Topics include:

- Innovative practices in data search used for journalism.
- Relevant technology (AI) in the media and creative industries more broadly.
- Data-driven media production (including its relationship to strong storytelling).

For the February 5th deadline, we would like to receive the following from each Partner. **Please use as much space as necessary.*

PART I: Needs Analysis
1. Self-assessment of the needs in skills and knowledge in data-driven technologies used by the media and other creative business professionals. Please, share your draft materials, which you did prepare while writing the MediaNumeric proposal. *Free form*.

Response:

2. Suggested list of questions for the semi-structured interviews about the needs in knowledge and skills with experts and ”average” professionals. Please, note what kind of respondent which question should be addressed:

Interview questions for the NEEDS ANALYSIS addressed to the national and international experts in:

the field of media and data *(incl journalism, marketing, big data)*, creative industries, incl marketing, heritage and the arts

Policymakers media, creative industries

Interview questions for the NEEDS ANALYSIS addressed to the professionals in higher education in relevant fields (specifically in degree courses for journalism, media and creative industries professionals)

If helpful to you, please fill in your questions in the chart below.
Please, feel free to add any relevant categories of respondents and topics:

	Type of Respondent:	Questions:
Data-bases		
Journalism		
Higher education (creative business etc.) - Investigative journalism - Fact-checking; - Data-visualisation etc.		

Interview questions for (young) professionals and (advanced) students in courses in:

	Questions:
News reporting	

(Investigative) journalism	
Local journalism	
Communications	
Creative business	

Arts and heritage	

3. Suggested questions for the open answer survey in France, Poland and the Netherlands, also including any and all international contacts who might have additional relevant insights, ideas and information. This is a three-part open question survey for educational needs in relation to the use of data for media, journalism and creative industries courses in higher education (universities and polytechnics). Respondents for this survey are national experts in France, Poland and the Netherlands (and additional international experts) are:

Experts in the fields of journalism, media production, media literacy, creative industries, events industries, data analysis; academically and professionally
Educational professionals

Please, consider the following type of knowledge and skills:

Critical skills - understanding the media landscape and the media and creative industries with a focus on the role and importance of big data (democracy and inclusion, business and economy, culture)

What *basic philosophical insights and concepts* are needed to ground critical work with data for media production, journalism and work in the creative industries?
What *critical knowledge of the media and creative industries* is needed specifically in relation to working with data?
What *media literacy* is needed in what regard?
What basic knowledge of national and international *legal restrictions and regulations?*

Response:

Practical skills - for media and creative industries production that focus on how data can improve quality (for democracy and inclusion, business, and for culture and aesthetics)

What user skills are needed for critical *reading and assessing* of available data
What skills are needed in *finding archives and other data suppliers*
What skills are needed for *gathering data*
What skills are needed for *basic manipulation of (clean) data sets*
What skills are needed specifically for *investigative and news journalism*
What skills are needed for (sustainable and responsible) *marketing and advertising*
What skills are needed *for creative production in a variety of fields* (events, festivals, artist management, television etc)
What *advanced skills for specialised data-based support* for other creative industry professionals are needed

Response:

Integrative skills in using data analysis critically for media and creative production (from being able to critically read an analysis of data reports, to process data to visualise and present [use of data analysis e.g. in storytelling for different types of mediated content]).

What competencies are needed for visualisation

What competencies are needed for storytelling

What competencies are needed for *multi-platform presentation* (technically and UX)

Response:

PART II. State of the Art

1. Partner country suggestions in the following topics, plus also suggestions for international leaders:

Best practices, innovation, challenges in data-driven story-telling and misinformation debunking

Higher education establishments: who provides the best training?

Journalism: who leads the field in data-driven journalism?

Creative industries: which companies or groups are particularly innovative in using data sets?

Response:

Large data sets, excellent storytelling techniques, debunking, trends

Where are the big data sets? How can they be accessed?

Recommendation of literature to study.

Recommendation of experts to consult.

Response:

Technology

How is it used?

How can it be used for stronger storytelling in journalism and the broader creative industries?

Who are the experts in your respective countries?

2. Suggestions of interview questions to put to the experts in the field.

Response:

3. Suggestions of survey questions to put to experts in the field.

Response:

Appendix IV: Summary of Partners' Input

Topics list: Per theme, specified for professional field and type of respondent
The following questions can be included in the interviews where and when you feel they are useful

	Professional Field	Type of Respondent	Questions & Topics
Data gathering: search and exploration multimedia data	Data-bases in GLAM (galleries, libraries, archives, and museums)	Experts and professionals in heritage and the arts: GLAM sector and historical institutions	How can big data inspire artists? How big data and access to them can help in historical research/history of art/etc? How do you find the quality and accessibility of data produced by the GLAM sector? Are digital archives and data that define them a source of inspiration for you? How can data be a source of inspiration for artists? How can the use of data be useful in a heritage organisation (or a media company) to better manage the collections, enrich them and editorialise them?
		Data engineer/ Chief data officer	What skills are required today in order to work with data?
		Big Data experts, archives curators, public data experts	What are the most efficient tools to use in data exploration and analysis? Are there any (data exploration) tools that require no or little programming skills? Good practices examples
	Journalism	News journalist	How do you usually gather your data/do your research?
		Investigative	What are the main sources

		journalist	investigative journalists rely on in their daily work? How can the data be useful/helpful for a journalist/investigator, to better analyse information? ICIJ: https://www.ICIJ.org/ https://www.ICIJ.org/inside-ICIJ/2021/02/ICIJs-datashare-platform-to-keep-growing-with-new-focus-on-collaboration/ What are the skills and knowledge in data-driven technologies most necessary in this type of work?
		Local journalist	What sources do you usually use when you need specific data? Do you use AI tools/data in your day-to-day activity? If yes, how? Cooperation with local communities, remote tools, specific problems, knowledge of legal obligations of local authorities (i.e. public data access)
		Editor-in-chief	Did you implement a dedicated unit for data analysis/data innovation? If yes, after X years of activity, what view do you have on the work of this unit and how do you analyse the cost-benefit ratio? On which criteria do you evaluate this work? Which is the interest for an editorial team to invest in a unit dedicated to data analysis? How has the issue of data-processing evolved in the last 10 years?

			What view do they have regarding the complexity of data processing? Which tools are used for data processing? Did they turn more complex or on the contrary easier? What choices in terms of human resources did you make (skill improvement of journalists from the editorial team or recruitment of experts in data/geeks)? How do you see the development of data-driven journalism in a 5-10 years horizon? According to you, behind the issue of data journalism, what is today the stake of data governance? Is the question of data in itself more interesting than the simple question of data journalism? How can the data be used to generate turnover for the (media) companies (i.e monetisation)?
		Policymakers	For what reason is it crucial today for a (media /creative) company to have a global data-driven strategy? How can a (media/creative) company manage data efficiently in a global approach?
	Communication and media makers	Communication officer Content creators Editors & directors	What are the primary sources and references presented when communicating to the readers? How are digital data used and checked?

			Are digital archives sexy?
	Other creative industries (festivals, booking agencies, documentary production)	Creative business professional	List three elements that would inspire you to work more with archival sources and big data. How can the creative sector help journalists acquire verified and safe data? Namely, the data generated by this sector? What kind of data does your company generate? How do you use it? Do you have a data scientist on your staff? What kind of database does your company use? How can the use of data be useful to reach a bigger audience and maximise ad revenue? To what extent the use of data can foster creativity?
		All kinds of media and creative professionals	Trends, good practices Tools, skills Anxieties, dilemmas
Telling stories with multimedia data	Journalism	News journalism, investigative and local journalism	Are you interested in new interactive storytelling formats based on multimedia? How is the audience interested in the content based on the data processing? Is there proven evidence of such an interest?
		Editors-in-chief	Does data journalism contribute to changing the image of the media? If yes, how?

			We talk a lot about data for editorial content but shouldn't we open the discussion more broadly on the issue of data for media audiences/advertising strategies? To what extent does the management of data within a media exceed the question of data journalism? What are the strategies of media companies regarding data-visualisation? Is it more a "showcase" or else a true strategy truly rooted in the newsrooms? Business models, strategies, economic perspective Users behaviour changes (research data if available)
		Arts & Heritage	New career models, the gap between traditional "arts" approach and more technology-oriented creation
	Communication	All types of communication specialists/experts	How can AI be useful for companies in their communication and marketing strategies, and how the use of AI tools/data can generate revenues for a company?
Tracking and Debunking Misinformation	Journalism	News journalist	How do you define the place of ethics in the current setup of news reporting? In 2021, what share of the public interest in media goes to data journalism?
		Investigative Journalism	How can investigative journalists fight mis- and disinformation?
		Local Journalism	How does the use of data journalism participate in the conquest of a wider readership, younger and increasingly

			connected?
	Higher education	Policymakers, Teachers, Researchers, Marketing specialists	How can we ensure that the students have a global vision of data, and especially the data-governance ecosystem, at the international level? The academic market needs analysis and curricula changes New curricula elements Formal/legal academic limitations Good practices in teaching forms/curricula
		Policymakers	How can we overcome the issues connected with GAFA influence on journalism? How/should we make journalist/media studies more aware of ethical challenges and problems in modern journalism?

Example Questions for the Experts' Interview:

Topic I. Gathering data

Working with large databases

1. What kind of digital data sources could be used for data-gathering in your professional field?
2. What one 'must' know about contemporary data sets (sources)?
3. In data science, it is important to formulate desired insides before the research starts. How important is this skill in your professional field? How can one learn to formulate a good "data question"?
4. How can large multimedia data sources, such as archives, be used for creating stories and news?
5. What skills one has to obtain to be able to find specific information in a data set?
6. What should one learn *to see in* a data set? What relationships and correlations between different features (columns, descriptions, labels etc.) are there?
7. What data exploration tools require no or little programming skills?

8. For the practitioners, please, consider the following questions: whether/ to what extent the interviewees are already working with data - what is their approach? What are the benefits and challenges?

Topic II. Verifying data

Tracking and Debunking Misinformation

1. What are the main issues of current information ethics? (think about ignorance, missing information, misinformation, disinformation and other forms of deception or incompetence).
2. What are the technologies to track misinformation?
3. What open-source intelligence tools are useful for tracking and checking information?

Topic III. Presentation of data

Telling stories with data

1. Where would you recommend starting learning data visualisation skills?
2. What type of data visualisation techniques can be used in your professional field? (images, charts, diagrams, animations etc.)
3. How can data visualisation help an author to present a story in greater depth or detail?
4. What students should know about automated journalism?
5. What are the ways (principles) of cross-platform content presentation?

Topic IV. State of the Art

Data journalism and multimedia storytelling with data

1. How has the emergence and availability of large data sets changed journalism and the wider world of story-telling?
2. How vital is knowledge and manipulation of data to journalism and multimedia storytelling?
3. Does the use of these large data sets change the way you approach and tell a story?
4. What does the relatively recent discipline of data visualisation bring to our understanding of the world around us and the issues that we face today?
5. Can you cite examples of excellent data visualisation and explain why they worked?
6. Can you provide an example of data visualisation that did not work properly?
7. How has technology, automation changed the way journalists and multimedia storytellers interact with big data?
8. Where do you see data taking journalism and multimedia storytelling in the future?

Topic V. State of the Art

Tracking and debunking misinformation

1. How has the dissemination of misinformation changed over the past few years, e.g. the speed that it spreads, how it spreads, the aims behind the spread, the technology used to create misinformation?
2. How have the techniques for tracking and identifying misinformation evolved, e.g. use of deep fakes, manipulation of images, developing technology to counter these developments?
3. What are the techniques for debunking misinformation?
4. Are there topics that are particularly prone to misinformation?
5. Could you cite some cases of misinformation that were particularly impactful and provide details about how this information was debunked?
6. Have there been cases when it was not possible to debunk news that you fundamentally knew was false?
7. What is the impact on our societies of misinformation, is there a real threat to democracy?

Appendix V: Cover Letter for Interview

Stakeholder Board Formal Engagement Email

Subject:
Invitation to MediaNumeric Stakeholder Board

Dear **XXXX**
It is my pleasure to invite you to be on the **Stakeholder Board** for the **MediaNumeric** project. MediaNumeric is a 3-year (2021-2023) Erasmus+ European Union funded project that seeks to develop a new innovative training programme to educate the new generation of students in journalism and communication studies. In particular, MediaNumeric students are prepared with the theoretical know-how and skills needed to wade through and use (big) data, tell enriched (multimedia) stories, and track and debunk misinformation.

The project consortium is made up of leaders in academia, media and audiovisual archives from four EU Member States: The Netherlands Institute for Sound and Vision (NL), Inholland University of Applied Sciences (NL), National Film Archive - Audiovisual Institut (PL), University of Social Sciences and Humanities (PL), AFP Agence France Presse (FR), Institut national de l'audiovisuel (FR), Storytek (EE), and EUscreen Foundation (NL).

The Stakeholder Board is made up of approximately 30 experts involved in media and journalism and communication studies including broadcasters, media professionals and industry platforms, experts in media ethics and media literacy, open knowledge, big data, artificial intelligence and machine learning, from across Europe.

The Stakeholder Board members, of which we hope you will be one, are being brought together to support the project by lending their experience and expert knowledge in developing the training programme's curriculum and exploitation, making sure the course takes the most recent developments in the field into account.

The Stakeholder Board will be invited to:

- Contribute to the Updated Needs Analysis
- Contribute to the State of the Art of Data-Driven Journalism Analysis
- Reflect on training programme content
- Inform on training programme evaluation methodology and integrating feedback
- Contribute to exploitation planning

Were you to agree to collaborate with us this would mean participating in interviews/questionnaires approximately twice yearly around the topics outlined above and also engaging in a handful of meetings around particular tasks. Meetings over the 3 years include:

Mar/Apr 2021 (virtually)

- Contribute to the Updated Needs Analysis Report & State of the Art Analysis of Data-Driven Journalism Report by participating in an approx 60-minute interview

Apr 2022 (face-to-face if possible)

- Reflect on training programme content & exploitation planning

Apr 2023 (virtually)

- Discuss the training programme evaluation and impact assessment and integration of feedback

Aug 2023 (virtually)

- Review final Exploitation Plan

Does this sound like something of interest to you?
Your experience and expertise are invaluable to MediaNumeric and we do hope you join our Stakeholder Board.

If you can please let me know either way **by the beginning of next week (Tues 9 Mar)** that would be great. Please say if you have any questions - I'm more than happy to speak about this further or send through more information on the project.

Kindest,

WP2 Research Engagement Invitation

Dear **XXXX,**
With this email, I am reaching out to request your participation in the research we are conducting for the **MediaNumeric** project. MediaNumeric is a 3-year (2021-2023) Erasmus+ European Union funded project that seeks to develop a new innovative training programme to educate the new generation of students in journalism and communication studies. In particular, MediaNumeric students are prepared with the theoretical know-how and skills needed to wade through and use (big) data, tell enriched (multimedia) stories, and track and debunk misinformation.

The project consortium is made up of leaders in academia, media and audiovisual archives from four EU Member States: The Netherlands Institute for Sound and Vision (NL), Inholland University of Applied Sciences (NL), FINA National Film Archive - Audiovisual Institut (PL), SWPS University of Social Sciences and Humanities (PL), AFP Agence France Presse (FR), INA Institut national de l'audiovisuel (FR), Storytek (EE), and EUscreen Foundation (NL).

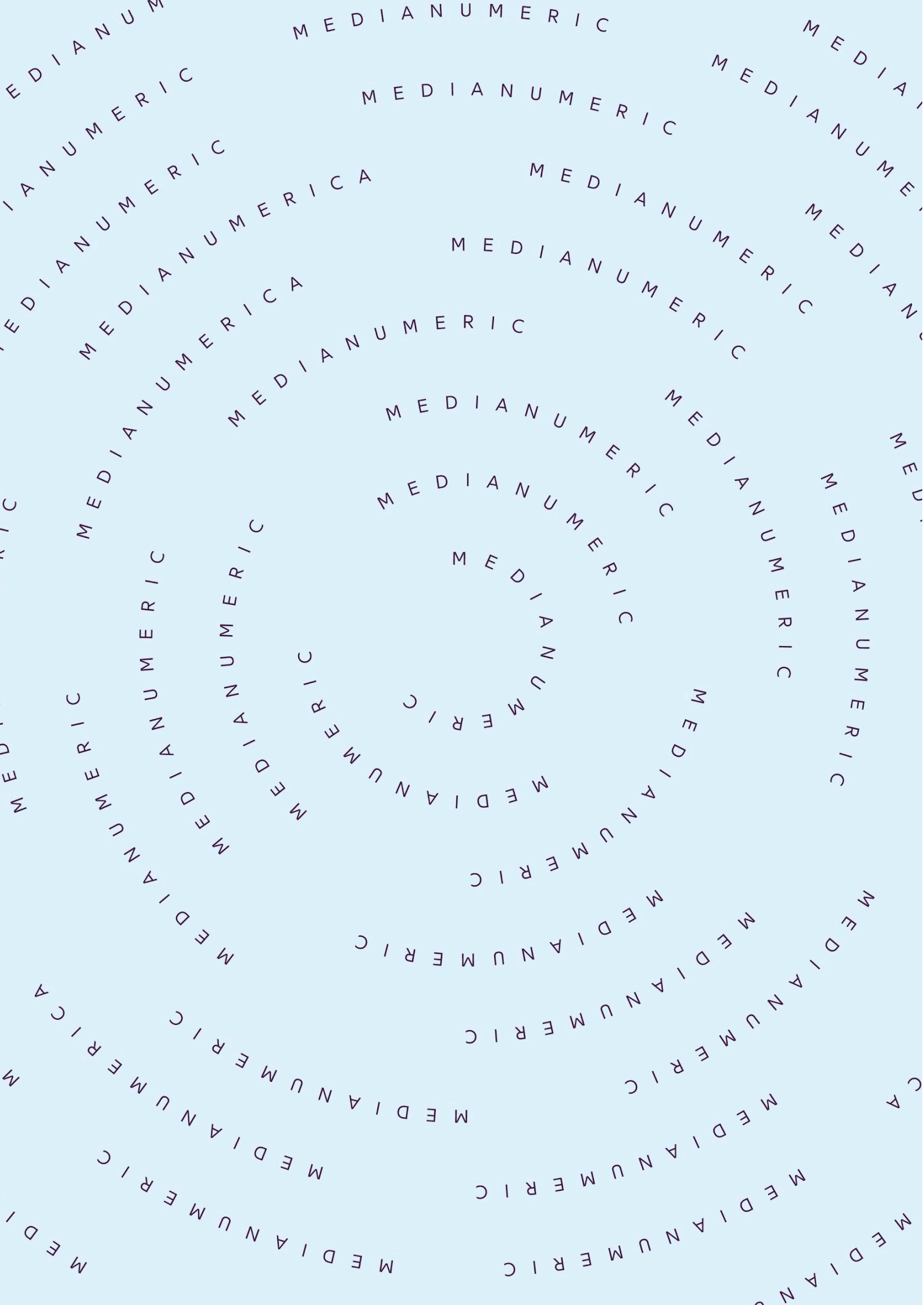
In particular, we are developing an Updated Needs Analysis and State of the Art of Data-Driven Journalism Analysis to guide the development of the training programme curriculum. Your experience and expertise would be an invaluable contribution here.

In particular, I'd like to plan an **approximately 60-minute interview** with you in the **next two weeks** to help generate content that will support the Analysis reports mentioned above.

I'm more than happy to share some sample questions ahead of time to help you prepare.
Is this something you would be interested in participating in?

If you can please let me know either way **by the beginning of next week (Tues 9 Mar)** that would be great. Please say if you have any questions - I'm more than happy to speak about this further or send through more information on the project.

Kindest,


MEDIANUMERIC