Lesson 1
Introduction to Data Journalism

Overview

A decade ago, data journalism was "very much a field 'in the making,'" according to Bounegru and Grey (2021). Now, it is "seen as a prominent label [that] refers to a diverse plurality of practices which can be empirically studied, specified and experimented with" (p. 12).

In this part of the course, we take the first few steps toward knowing data journalism—what it is, how it evolved through time, and what it can contribute to the journalism industry and to society.

Duration: 2 hours

Introduction

Through time and in many parts of the world, societies become digitized, information is stored in data formats, and, where freedom of information is recognized, the public is able to get hold of data they need for a variety of purposes (Knight, 2015). Data, according to Rogers, Schwabish, and Bowers (2017) (1) helps reduce complexity and thus enables readers to make sense of the world around them; (2) keeps society rooted in facts; and (3) improves newsrooms. The increasing availability and accessibility of data, coupled with continuing technological advances, present opportunities for newsrooms to glean insights on issues that impact on governments, communities, and people at the international, regional, national, and local settings. Data and algorithms are now additional inputs to journalism, which is a product of artistic/creative storytelling and a scientific approach to research.

Objectives

At the end of this lesson, you should be able to:

1. Define data journalism;
2. Describe the nature and features of data;
3. Discuss the potentials of data journalism in the development and communication landscape

Lesson proper

Discussion flow:

1. Definition of data journalism
2. Historical development of data journalism
3. Nature, features, and potentials of data journalism

DEFINING DATA JOURNALISM

What is data journalism? What do we mean by *data*? Does journalism come before data, or is it the other way around? What makes data journalism different from the other forms of data-driven journalism? These are a few of the usual questions we ask when we first encounter the term—questions that we will attempt to answer in this lesson.

The term *data journalism* includes the words *data* and *journalism*, both "troublesome terms" according to Bradshaw (2012, p. 2). The first of the two would usually connote numbers and spreadsheets, which can cause people to automatically refuse dealing with it owing to unpleasant experiences or lacking the requisite skill. The latter is being challenged from many sides amid issues of credibility and relevance. Data journalism is "a deeply contested and simultaneously diffuse term" (Flink & Anderson, 2015 p. 468), the heart of the matter being what it actually is and what it encompasses (Knight, 2015). Lewis (2021) writes that disagreements about data journalism are on matters such as "whether numbers must be central to the reporting, how large the data set should be, and whether visualizations are necessary or sufficient to qualify as data journalism" (p. 79). No consensus definition exists at the moment, and much of what is written about data journalism is based on the varying experiences of practitioners and ponderings of scholars.

Through the years, various authors forwarded definitions of *data journalism*, including descriptions that pertain to both content and form of data journalism. One of the simplest is that written by Bradshaw (2012): data journalism being simply journalism done with data. A note: Data per se has been around for a long time, even long before journalism was born; centuries ago, there was an accounting of deaths and births in cities. For our purposes, by *data* we mean *big data*, the kind that is voluminous and thus necessitating the use of computers and software to be gathered, processed, and interpreted. Aside from volume, big data is characterized by its variety (presence of complex data coming from different sources), veracity (accuracy and integrity of data), and velocity (speed of gathering data). In the same vein, data journalism is journalism that uses big data as a story source. Now

onto the next term needing clarification—*use*, i.e., what is meant by *using (big) data.* In this regard, we may find the other definitions more insightful as these include the various ways by which data is used in journalism. Other definitions emphasize the purpose for which data is used, i.e., the end goal of data journalism. Several of such definitions are found in the table below.

Emphasis	Definition	Author
Elements	"a field combining spreadsheets, graphics data analysis and the biggest news stories"	Rogers (2011, cited in Knight, 2015 p. 58)
	"the inheritor of two older news practices: [news] infographics and computer-assisted reporting"	Knight, 2015
Processes	"gathering, cleaning, organizing, analyzing, visualizing, and publishing data to support the creation of acts of journalism"	Howard (2014 p. 4)
Purpose	"using data to tell stories in the best possible way, combining the best techniques of journalism"	Rogers (2016, cited in Lewis, 2021 p. 80)
	"the practice of finding stories in numbers and using numbers to tell stories"	Broussard (quoted in Howard, 2014 p. 5)

What should be clear to us at this point is that doing data journalism entails (1) working with data, (2) finding a story in it, and (3) delivering such a story using different tools and techniques. An important point to always remember is that notwithstanding the volume or the complexity of data that we may be dealing with, data journalism *is* journalism—this means that the emphasis is on the journalism and not on the data, which is treated as a component of journalism harnessed to tell a compelling (and often complex) story that a journalist with a nose for news would find. Further, that (big) data is used in journalism means that the usual breaking the story may take a back seat vis-à-vis "telling us what a certain development might actually mean" (Lorenz, 2012 p. 4).

According to Lewis (2021), definitions of data journalism are fluid owing to the reliance of this practice on technology. As mentioned earlier, no consensus definition is currently available. Can you offer a definition of data journalism on the basis of the above discussion and your knowledge of technology or tools that journalists can use in reporting?

==========

Learning activities

Watch Video recording of Day 1 of the 2020 DDJ Workshop Series: Data Journalism, featuring Karol Ilagan of the PCIJ or the Philippine Center for Investigative Journalism. Available at https://tinyurl.com/DDJworkshopdatajourn080320

Read Chapter 1 of *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News* edited by Grey, Bounegru, and Chambers

==========

HISTORICAL DEVELOPMENT OF DATA JOURNALISM

Pinpointing the exact moment when data journalism was born is challenging. In the same way that numerous definitions have been proposed toward understanding what data journalism actually is, different historical highlights are being deemed as forerunners of present-day data journalism.

The use of computers and databases for storytelling is the current manifestation of data journalism, although use of data in reporting dates back to more than a century ago (Anderson, 2018). Knight (2015) calls data journalism as "the inheritor of two older news practices: [news] infographics and computer-assisted reporting" (p. 56), the first practice being the production of factual illustrations, such as graphs, charts, and maps, dating back from the late 19th century, while the second being the use of computers to gather and analyze data to be used in news stories, was widely used in the 1980s.

Let us look at several key moments in the history of using data in journalism in order to trace the development of what Anderson (2018) calls "quantitative truth building" (page 1).

Historical period	Events / Highlights
1821	As Rogers (2013) claims, the *Guardian* was already practicing data journalism since its foundation year. A table of information supposed to have been leaked to the paper's editor showed the schools in Manchester, the number of students who attended, and the costs per school, thereby

	helping reveal the actual number of students receiving free education 70 years before education became compulsory (Bounegru, 2012; Knight, 2015).
1842	In September of this year, *The New York Tribune* published on its front page a line chart that showed the number of deaths in New York City from a cholera epidemic. The chart is described as "a snapshot of the state of the art of data visualization in news at that moment and is full of clues that help reveal parts of the hidden history of visual journalism" (Klein, 2016).
1858	Florence Nightingale wrote a report entitled "Mortality of the British Army," wherein she used graphics to advocate improvements in health services. Part of the report showed that the majority of deaths in the army were from preventable diseases and not bullets (Bounegru, 2012).
Early 20th century	Social survey reportage
1950s	Computer-assisted reporting (CAR), the first organized and systematic approach to using computers for data gathering and analysis for journalism, was first used in 1952 by the CBS to predict the result of the presidential election (Bounegru, 2012).
1960s	US-based investigative journalists started analyzing databases of public records using scientific methods in their effort to monitor power. Their journalism was known as "public service journalism" because their work "sought to reveal trends, debunk popular knowledge, and reveal injustices perpetrated by public authorities and private corporations" (Bounegru, 2012, p. 18).
1970s	The term *precision journalism* was coined to describe "the application of social and behavioral social science research methods to the practice of journalism" and as a response to what was then called *new journalism* or the application of fiction techniques to journalistic reporting; precision journalism was used "to represent marginal groups and their stories" (Bounegru, 2012, p. 19).
1980s	Launched in 1982, *USA Today* was said to have revolutionized the newspaper graphic with its colors, maps, and boldness; later it was criticized for "dumbing down the news [by] reducing information to diagrams and pictures" (Knight, 2015 p. 56).
1980s	The growth of personal computers, the internet, and expertise in computing contributed to the wide use of CAR. In 1989 the National Institute for Computer-Assisted Reporting was founded in Missouri, USA (Knight, 2015).
1990s	"The New Precision Journalism" and other textbooks on CAR were published and served to bring the idea of CAR into mainstream journalism and journalism education (Knight, 2015).
Early 21st century	Considerable research was done about introducing CAR into the journalism curriculum but less on the actual use of CAR in newsrooms. There was notable increase in the use of graphics in reporting, and "the availability of

	data and access to the means to analyze it continued through the first decade of the 21st century" (Knight, 2015). In 2006, Holovaty wrote one of "the earliest formulations of data journalism"—an essay arguing the need for newspaper sites to change, specifically for journalists to publish structured, machine-readable data along with the traditional article. By *structured data* Holovaty means "information with attributes that are consistent across a domain" (Bounegru, 2012 p. 18). In 2009, the Guardian Data Blog was launched. It is described as "the first systematic effort to incorporate publicly available data sources into news reporting" (www.stateofopendata.od4d.net).
2010s	The start of the decade saw the start of conversations about opening government data to the public by publishing it online. Around this time, analyzing data was deemed "the future of journalism." In July 2010, the *Guardian* began publishing data journalism about the leaked Afghanistan war records, called the "War Logs." In the years that followed, data journalism began catching fire (Howard, 2014). Data journalism teams and initiatives started democratizing data through creating and using tools, such as apps (case of the Investigative Dashboard of the Organized Crime and Corruption Reporting Project), launching a data store (ProPublica in 2014), creating data journalism teams or units (such as La Nacion of Argentina in 2010), and releasing reports (www.stateofopendata.od4d.net)

Taking off from the above discussion, where do you think data journalism is headed in the short, medium, and long term?

==========

Learning activities

Watch Video recording of "History of data journalism at the Guardian, available at https://www.theguardian.com/news/datablog/video/2013/apr/04/history-of-data-journalism-video and at https://youtu.be/iIa5EoxyvZI

Read Online articles about using data in journalism:

- "Fifty Years of Journalism and Data: A Brief History" by the Global Investigative Journalism Network, available at https://gijn.org/2015/fifty-years-of-journalism-and-data-a-brief-history/
- "Infographics in the Time of Cholera" available at https://www.propublica.org/nerds/infographics-in-the-time-of-cholera

==========

Nature, Features, And Potentials Of Data Journalism

An overwhelming quantity of data is currently available to us, but the sheer volume can readily discourage anyone from taking a look to understand it and glean insights. What Meyer (2011) describes as "an endless stream of data" needs explaining or can enrich storytelling, and such are part of what data journalists do. Rogers, Schwabish, and Bowers (2017) identify three types of stories produced from data journalism, dubbed by Wright and Doyle (2018) "forms of data journalism," namely: stories enriched by data, stories that use data to investigate, and stories that explain data.

Stories enriched by data. These are our traditional news stories wherein relevant data is used to support reporting. Journalists use data to enrich or illustrate stories and to provide evidence for a story point of view. This kind of data journalism is not fundamental to a story, i.e., the story can exist without it, "but the data rather enriches and fortifies it" (Rogers, Schwabish, & Bowers, 2017 p. 14). Wright and Doyle (2018) describe this form of journalism as "smaller-scale, everyday data journalism" (p. 2). Examples of this work include:

- "The Strongest Evidence Yet That America is Botching Coronavirus Testing" by Robinson Meyer and Alexis C. Madrigal, available at https://www.theatlantic.com/health/archive/2020/03/how-many-americans-have-been-tested-coronavirus/607597
- "To use or not to use: How the govt's face shield policy evolved" by Vera Files: https://verafiles.org/articles/vera-files-fact-check-use-or-not-use-how-govts-face-shield-p

Stories that use data to investigate. These may be deemed as "investigative data journalism" (Rogers, Schwabish, & Bowers, 2017 p. 13), in which stories are uncovered through data. Journalists expose information or surface a story hidden in data. Such stories rely on a combination of skillsets, may involve large teams, require more resources, and take longer to produce. Examples of this work include:

- "Land-grab universities" by Robert Lee and Tristan Ahtone, available at https://www.hcn.org/issues/52.4/indigenous-affairs-education-land-grab-universities
- "For banks that backed PH coal boom, the path to renewable energy comes with roadblocks" by Karol Ilagan, available at https://pcij.org/article/6625/for-banks-that-backed-ph-coal-boom-the-path-to-renewable-energy-comes-with-roadblocks

Stories that explain data. As data becomes increasingly available and accessible to the public, explanations and context are necessary so that it becomes comprehensible and relevant to audiences. Journalists thus act as bridge or guide "between those in power who have the data—and are rubbish at explaining it—and the public who desperately want to understand the data and access it but need help" (Rogers in Rogers, Schwabish, & Bowers, 2017 p. 15). Examples of this work include:

- "Dollars for Docs: How Industry Dollars Reached Your Doctors" by Mike Tigas, Ryan Grochowski, Charles Ornstein, and Lena Groeger, available at https://projects.propublica.org/docdollars
- "Voter Statistics and Elected Candidates" by Crystal Joy De La Rosa, Justin Oliver Fiestada, Faye Gali, Tiny Tayam, R-Jay Sale, and Lloyd Macalalad, available at https://moneypolitics.pcij.org/#/story/Voter-Statistics-and-Elected-Candidates/

Howard (2014) says that data journalism has the following components:

1. Treatment of data as source to be gathered and validated;
2. Application of statistics to interrogate data; and
3. Visualization to present the data

Treatment of data as source to be gathered and validated. Stories always come from somewhere or someone: a newsmaker, an asset, an anonymous tip, a data set, etc. Data is a story source that can complement experts, documents, narratives, and other story inputs. Rarely does it fall on a journalist's lap ready for use—just like other sources, oftentimes it needs to be sought and verified before being used.

Because journalism is in the business of truth telling, the accuracy, completeness, and veracity of inputs to a report need to be ensured. Evaluating the trustworthiness of sources is a must, be these sources a person, a pile of documents, or datasets. In assessing the trustworthiness of big datasets, a journalist should also understand the big picture of the process and infrastructure that enables the collection of big data. Datasets do not exist in a vacuum—its context can be explained by metadata, its collection, and the process it undergoes on transformation and storage.

Application of statistics to interrogate data. As a human source would be interrogated by a journalist, so should data be toward gaining insights on the quality of the data, finding answers to important questions, spotting trends, or identifying gaps that will present opportunities for improving the dataset. Toward this end, various tests may be employed. Remember that a suitable test (or tests) need to be used lest we obtain misleading results.

Visualization to present the data. In many instances, the story is best told using visuals, such as charts and graphs. Such a presentation would usually be enhanced by interactive features that make data storytelling not just informative but also interesting. A visual representation also aids in summarizing the numbers that seem to be daunting viewed in thousands of rows.

Key areas of data journalism

According to Howard (2014), four key areas comprise data journalism, namely, data reporting, data visualization and interactives, emerging journalistic technologies, and computational journalism. The table below gives details about each of the four categories, as presented in *The Art and Science of Data-driven Journalism*:

Categories	Inclusions	Techniques and technologies
Data reporting Obtaining, cleaning, and analyzing data for use in telling journalistic stories	• Deploying computer-assisted reporting or analysis for writing journalistic stories; • Practicing precision journalism, including the use of social science research methods in the interest of journalism; • Visualizing data for use in exploration and analysis; and • Programming to obtain and analyze data for writing journalistic stories	• Invoking public records law to negotiate for data; • Using web scraping tools and techniques; • Using relational database software; • Understanding statistical concepts and software or programming languages with statistical packages; and • Using mapping and visualization tools and software
Data visualization and interactives	• Visualizations developed and designed as interactive charts and	• The use of code; • The use of visualization software or programs;

Using code for digital publishing, as well as programming and database management to build interactive journalistic work	graphics for presentation; • Interactive applications, including searchable databases and games that help readers explore and understand a news story	• Database management and programming; • Mapping applications; • Server knowledge and the use of GitHub, versioning, and Agile software development technique
Emerging journalistic technologies New developments using data and technology	• Drone journalism • Sensor journalism • Virtual and augmented reality journalism,	• Drone technologies, such as an airframe, an autopilot of varying capabilities, a control system, and a sensor • Sensor technologies include a wide range of software and hardware to measure physical conditions; these are used to gather data with a small, portable computer or microcontroller • Virtual and augmented reality technologies are helpful in facilitating interactivity, creating scenes by stitching together camera outputs, etc.; however, questions of narrative, audience interaction, and journalistic values have yet to be settled with these technologies
Computational journalism The use of algorithms, machine learning, and other new methods to accomplish journalistic goals; this area overlaps with data reporting and emerging technologies	• Algorithms that help journalists mine unstructured data in new ways • New digital platforms to better manage documents and data	• Programming languages and applications that enable journalists to mix code and prose as they perform analysis and show the steps in their work; • Platforms that facilitate the use of complicated computational processes, like natural

	language processing and topic modeling

Potentials of data journalism

Use of data in reporting, according to Anderson (2021) is "inevitably intertwined with national politics, the evolution of computable databases and the history of professional scientific fields" (p. 353). Data journalism cannot be conveniently divorced from technological innovations that democratized access to and analysis of big data, as well as the quantitative approaches of social sciences that help to problematize and make sense of social phenomena.

Creating new roles for journalists. Data may be available and accessible, but it is not necessarily understandable—hence the need for journalists to help the public understand what it means and its implications to society. Using new and various approaches to storytelling developed from or alongside advances in technology, data journalists serve as *intermediaries* between the data and the audience by turning complex data into digestible and meaningful stories. As mentioned earlier, journalists also *bridge* the people with the data and the public that needs it (Kalatzi, Bratsas, & Veglis, 2018 p. 39 citing Gray et al., 2012, Rogers, 2013 and Lesage & Hackett, 2014). Further, data journalists help people make sense of data and issues so they can avoid confusion and misunderstanding.

Strengthening watchdog function of journalism. Data journalists publish datasets and reports in "a way that the public can control those in power" (Hamilton et al., 2009 in Kalatzi, Bratsas, & Veglis, 2018)—which include not just people in government but also those in newsrooms who have an obligation to do transparent and honest reporting. Numbers, when properly contextualized and analyzed, can provide tangible proof and be used to hold into account people and organizations entrusted with the care of a government and to also speak truth into power.

Presenting a new facet of journalism. Data journalism opens avenues for collaboration among professions whose knowledge of and work with data can contribute to journalistic storytelling, such as data scientists, programmers, and statisticians. It also allows for audience engagement with data through interactive elements, such as visualizations. Aside from using big data to tell a story that are already mainstream interests, data journalism

also opens new opportunities for stories such as investigating algorithms that govern our daily lives.

Data journalism is a relatively new form of storytelling and reporting, and perhaps it is this newness that accounts for the many interesting opportunities it opens to newsrooms. While challenges and limitations are present, the good outweighs the not-so-good, as can be seen in the data journalism projects that earn international recognition.

Anderson (2021) writes of data journalism:

> *Data journalism may be the most powerful form of collective journalistic sense making in the world today. At the very least, it may be the most positive and positivistic form of journalism. This power (the capacity of data journalism to create high-quality journalism, along with the rhetorical force of the data journalism model), positivity (most data journalists have high hopes for the future of their particular subfield, convinced it is on the rise) and positivism (data reporters are strong believers in the ability of method-guided research to capture real and provable facts about the world) create what I would call an empirically self-assured profession. One consequence of this self-assurance, I would argue, is that it can also create a Whiggish assumption that data journalism is always improving and improving the world. Such an attitude can lead to arrogance and a lack of critical self-reflexivity, and make journalism more like the institutions it spends its time calling to account.* (pp. 351-352)

In line with the above, how do you see data journalism a decade from now? What challenges should have been overcome and what opportunities should have been explored by then? What can account for such accomplishments of data journalism, if these can be deemed as such?

==========

Learning activities

Read "Genealogies of Data Journalism" by C. W. Anderson in Bounegru, L. and J. Gray (eds.), *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam: Amsterdam University Press. DOI: 10.5117/9789462989511_ch48

Do Look up a few of the data journalism projects referenced in this lesson. Revisit each and ask yourself the following questions:

1. What are the characteristics of good data journalism? How about bad data journalism?
2. Across time, how did data journalism change vis-a-vis available technologies and social issues?

==========

REFERENCES

Anderson, C. W. (2018). *Apostles of certainty: Data journalism and the politics of doubt*. Oxford University Press.

Anderson, C. W. (2021). Genealogies of Data Journalism. In Bounegru, L. and J. Gray (eds.), *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam: Amsterdam University Press. DOI: 10.5117/9789462989511_ch48

Davies, T., Walker, S., Rubinstein, M., & Perini, F. (Eds.). (2019). *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and International Development Research Centre.

Fink, K. & Anderson, C. W. (2015). Data Journalism in the United States, *Journalism Studies*, 16 (4): 467-481, DOI: 10.1080/1461670X.2014.939852

Four Key Areas of Data Journalism · Teaching Data and Computational Journalism (gitbooks.io)

Gray, J., Chambers, L., & Bounegru, L. (Eds.). (2012). The data journalism handbook. First edition. Sebastopol, CA: O'Reilly Media. http://datajournalismhandbook.org/

"A fundamental way newspapers need to change." Retrieved from http://www.holovaty.com/writing/fundamental-change/

Howard, A. B. (2014). The art and science of data-driven journalism. New York, NY.

Kalatzi, O., Bratsas, C., & Veglis, A. (2018). The Principles, Features, and Techniques of Data Journalism, *Studies in Media and Communication*, 6(2): 36-44. DOI: 10.11114/smc.v6i2.3208

Klein, S. (2016). "Infographics in the time of cholera." Retrieved from https://www.propublica.org/nerds/infographics-in-the-time-of-cholera

Rogers, S., Schwabish, J. & Bowers, D. (2017). "Data Journalism in 2017: The Current State and Challenges Facing the Field Today." Accessed 30 June 2021 https://newslab.withgoogle.com/assets/docs/data-journalism-in-2017.pdf.

Tow Center for Digital Journalism Publications. Retrieved from http://towcenter.org/wp-content/uploads/2014/05/Tow-Center-Data-Driven-Journalism.pdf

SUMMARY OF LESSON 1

No consensus definition of data journalism exists. Current attempts at defining the term are based on the proponents' experience and appreciation of data journalism and its elements, processes, and purpose. In the same vein, the history of data journalism has shown how it has evolved from reporting of tables in the 1800s to more concrete examples in the early 70s with 'precision journalism' and 'computer-assisted reporting' in the early 90s to how it has taken shape with the aid of new technology and platforms in 2010. The importance of data journalism has been highlighted as it creates new roles while strengthening existing functions of journalists.

SELF-ASSESSMENT QUESTIONS

After learning about data journalism, reflect on its place in development journalism practice. Answer the following questions in Canvas (this is not a graded output):

1. How do you see data journalism vis-à-vis development?
2. What opportunities do you see opening up for data journalism to contribute to conversations about development issues?