UNIT 8 UNDERSTANDING DATA JOURNALISM

Structure

8.0 INTRODUCTION

With the transformation of the brick and mortar economy into a digital economy, there has been a change in the way journalism is done. Data journalism is among the most significant changes visible and inevitable today. Data is described, arguably, as the 'new oil' with the connotation that whoever controls this precious resource will control the world in the coming years. Although the use of data, surveys and computer-assisted reporting (CAR) for journalism has been around for decades, journalists have started using new technologies to access seriously large amounts of data relatively recently. Today, the new and emerging technologies are helping both writers and readers to "understand data, explore it, and act upon the insights derived from it"[1] to make sense of the world much better than ever before. Data

journalism, as we understand it today, takes journalists and their audiences from the mere numeracy to visualisation of the larger picture and beyond, through a large-scale knowledge mobilisation, and an understanding of socio-economic trends and patters.

8.1 LEARNING OUTCOMES

At the end of the course, students should be able to:

- explain the meaning and importance of data and its connotations in the digital world such as public domain, privacy and data protection.
- familiarise themselves with the power of data in a country's development and also in various democratic, colonial and neoliberal contexts. Use time-series data about budgets, census, economic growth, human development, education and health.
- describe concepts like Algorithms, Artificial Intelligence (AI), computerised aggregation and data surveillance
- combine the ability to use data with narratives and storytelling techniques and think critically about the strengths and weaknesses of data use.
- analyse and visualise data (including spreadsheets and computer other assisted programmes) by using software packages and applications
- assess how governments, industry and different institutions may be collecting and deploying time-series data and their use in public life
- describe the debate around open source data and unauthorised leaks with or without the public interest

8.2 HISTORICAL BACKGROUND

The newsrooms of the eighties and the nineties assigned statistical analysis to business or research departments but data analysis as a tool for basic journalism was missing. Prior to data journalism, computer-assisted reporting (CAR) existed since the 1950s. CAR used relatively modest amounts of data for analysis and understanding patterns with the help of database software.[2] In 1952, American news network CBS used CAR for the first time to predict the result of the US Presidential elections. The US investigative journalists since the 1960s started analysing databases of public records with scientific methods so as to create their own stories.[3] Investigations carried out by Phil Meyer of the Washington Bureau of the Free Press and his team used data to find that the Detroit riots were prompted by lack of jobs, poor housing, crowded living conditions, and police brutality.[4] The use of social science survey methods coupled with CAR gave rise to precision journalism.[5]

8.2.1 The Indian Experience

In India the use of big data in journalism was historically through departments like the National Sample Survey Office (NSSO) under the Ministry of Statistics and Programme.

Implementation and the Census of India which has collected, collated and analysed big data every ten years since 1872. After Independence in 1947, the task which

was started by the colonial rulers was continued by the Registrar General and the Census Commissioner of India under the Ministry of Home Affairs. Successive Indian Governments have also followed the tradition of releasing extensive and reliable data about the state of the economy in the form of the Economic Survey just before the union budget. The Economic Survey and the presentation of the Union Budget in Parliament have always been sources of intense debates in the Indian media (Including the Railway budget until the 92-year old practice was discontinued after the two budgets were merged in September 2016). The Reserve Bank of India also releases data through its periodic and annual reports which are thoroughly analysed in the economic papers.

So, what has changed since then? Well, technology has made it a lot easier to use machine-readable data for more accurate and longer-term analysis. Also, earlier data analysis was

assigned strictly to a handful of specialists and everyday journalism rarely deployed large amounts of data for routine work, with the notable exception of elections. Among the first users of big election data in India were David Butler, Ashok Lahiri and Prannoy Roy (1989) who used the Election Commission data of 37 years (from 1952 to 1989) for a seminal book and for extensive analysis in the media.[6]

However, in today's newsrooms, a certain amount of familiarity with data analysis is required for all departments, including the desk, reporters, senior editors and even designers and illustrators. It is also common for modern newsrooms to conduct or outsource opinion surveys or demographic analysis with the use of specialised fields like psephology and urban geography with enormous amounts of data. Data journalism is fast emerging as a tool of investigating journalism, and a means for demanding transparency and accountability in governance. Modern mass media cannot fulfil its watchdog function without harnessing and analysing data. The whole issue of data for analyses of trends is also opening up new debates about privacy and data protection. A data journalist can tell a complex story quite simply and convincingly by using engaging and colourful infographics. The idea is to show with the help of data what the naked eye cannot see or what cannot be described or understood through quotes from usual sources of news stories.[7]

8.2.2 Data Surveillance: The Dark Side of the Internet

Internet is often called a democratic space or a real level playing field because it does not differentiate between the rich and the poor users. True that the digital divide puts poorest of poor people in the global South at a huge disadvantage but that does not take away the fact that the Internet does not differentiate (or discriminate) between individual users. However, it is increasingly being argued that the democratic nature of the Internet could well be, at best a myth and at worst a ruse, for promoting predatory market forces.

Yasha Levin (2018) argues in "Surveillance Valley: The Secret Military History of the Internet" that the Internet was in fact built by the powerful (US) government to spy on the citizens. The author believes that the manipulation of personal data of the unsuspecting users is quite central to the political economy of the powerful Western countries like the US. Looking at data from the perspective of the new and emerging tech-industry giants like the

Facebook, Google and Twitter and marketing companies like the Amazon, he argues that the business of these companies is to spy on their users by collecting a variety of data in order to maximise their own profits.[8] Bringing in the emergence of a new military-digital-complex the author asserts that the military and the industry often complement one another. Nick Couldry and Ulises A Mejias (2018) argue that the 'capture and processing' of data (social or personal) unfolds through new "data relations" which may be understood through new forms of data colonialism.[9] That is why most governments in all parts of the world are getting incrementally more circumspect about protecting the data of their individual citizens from predatory digital invaders.

8.2.3 Wiki leaks and the Gorilla war on Data Secrecy

The control over data has witnesses some pitched battles between people who want to store classified data, mainly governments and big multinational corporations, and those who want to declassify them in public interest. The data in high demand can be about illegal bank accounts maintained broad by the rich and powerful or data about crucial public information intentionally kept away from the citizens. The subject has serious repercussions for the future of journalism, data secrecy and the privacy of citizens. It also has strong ethical dimension because the unauthorized sharing of classified data may often be a crime in many countries. The opponents of wholesale data leaks believe it to be a form of theft, its supports say just like sting journalism, this form of data journalism may be justified against the public interest it serves. What works in favour of the citizens are a variety of anti-secrecy laws such as the Right to Information Act, Whistle-blower Protection Act or the Data Protection Act.

At the forefront of the war on secrecy is the WikiLeaks which is a global non-profit organisation dedicated to leaking classified information which it believes should be in public domain. Founded in 2006 by Julian Assange, the organisation promotes data journalism by unearthing and analysing large databases regarding corruption, unethical practices, illegal secret operations and covert wars etc. One of the biggest data leaks after Assage was done in 2013 by another computer scientist and a former employee of the Central Intelligence Agency (CIA) Edward Snowden. The hacker, who is often described as a traitor (because of unauthorised leaks), a dissident or a whistle-blower (because of the nature of information leaked, ostensibly in public interest), has disclosed numerous global surveillance programmes of the US Government and its covert, often questionable, foreign-policy operations.

According to the Wikileaks' website, it has published over 10 million documents and related analysis so far. WikiLeaks is intensely disliked by many powerful governments but it is also a recipient of some of the world's most coveted journalism awards. Der Spiegel, one of the most influential German publications, quotes Julian Assange where he describes Wikileaks as "… a giant library of the world's most persecuted documents. We give asylum to these documents, we analyse them, we promote them and we obtain more."[10] The organisation claims to have "contractual relationships and secure communications paths to more than 100 major media organisations from around the world. This gives WikiLeaks sources negotiating power, impact and technical protections that would otherwise be difficult or impossible to achieve."[11]

8.3 IMPORTANCE OF DATA JOURNALISM

According to a report from Tow Center for Digital Journalism, data journalism is the "application of data science to journalism, where data science is defined as the study of the extraction of knowledge from data".[12] In short, data journalism is about narrating stories or analysing social and political processes around us with the use of numbers or finding stories/ patterns in numbers. The use of humongous amounts of data has been made possible by computerisation, new and emerging technologies, digital storage devices, and the opportunities of easy sharing through the Internet and multiple public domains.

Due to the explosion of data which is generated by various sectors including government, wider society, industry, academia, individual citizens, social media and the research community, both the context and scope of data journalism have expanded in the recent years. One can observe that data journalists often use free, powerful online tools and open source software to collect, clean, and publish data in interactive features, mobile apps, and digital maps.[13] The availability of free and open source data in a convenient and modifiable form too has lent a helping hand to data journalists. Unlike in the past, when the global trends took a long time to catch up with the developing countries, data journalism is already an accepted phenomenon all over the world, including in India.

8.4 BASIC COURSE REQUIREMENTS OF DATA JOURNALISM

Data journalism needs systematic understanding of certain specific areas in the form of building blocks. If the students are comfortable with the idea of numeracy and develop basic skills of making sense of spreadsheet-based programmes, they will be able to use numbers for deeper analysis of things. The course will prepare students to learn the basic proficiency in cleaning data for their own understanding, analysing it for interpretation in the form of conclusions. In the end, they should be able to interpret quantitative academic research, tables rows and columns or from specialised programmes like Excel, spreadsheet programmes, Google Sheets and advanced visualization techniques.

The following steps can be seen as the building blocks of data journalism:

1. Gathering of data and distinguishing between reliable and unreliable sources of data. Developing an understanding of the public domain and techniques to access publicly available data.

2. Use of spreadsheets and various popular software programmes for digital number crunching. This is more of a skill which can be learnt in simple steps and with classroom-based practical experiments. As time passes, newer and better tools of data presentation and analysis will be available, including methods using artificial intelligence (AI) with the help of mobile phone Apps.

3. Visualising data as part of a story. Too many numbers in a story can burden the narrative and make it stodgy and unreadable. However, imaginative use of charts, tables and graphics can increase the depth and gravitas of the story. This would require a team work but the reporter must be able to first visualise a story in which data, narratives and even digital maps and pictures can be combined.

4. The Editors and senior journalists must be able to question 'data-ism' where data takes over human element in a news story. As a tool of empowerment, data should be able to enhance journalistic insights and intuitions rather than creating confusion. It is equally important for the user of data to understand how incorrect, insufficient data can create false paradigms and how to differentiate between good, bad, biased and motivated data.

Check Your Progress 1

Note: 1) Use the space given below for your answer
2) Compare your answer with those given at the end of this Unit.

1. How did Data journalism begin and what are the important changes it has undergone in the digital world?

.....

.....

.....

8.5 AGGREGATORS AND ALGORITHMS

In the light of the busy lifestyles, ever increasing dependence on the Internet, and focus on getting personalised information, Content Aggregators play a significant role as purveyors of news and information on the go. News Aggregators offer an automated platform which deploy search engine optimization tools to collate and disseminate the news that has been reported by other sources. The stories which are "trending" on top of aggregated platforms are considered to be good and many news organisations compete with one another to get the top slot on, say, Google News.

In terms of computing, news aggregators (also called feed aggregators) amass pieces of information in the form of text, audio-video clips and pictures etc and present them in friendly formats to existing and potential users in order to save their time and effort. To make the news more and more relevant to the readers, the aggregators often collect personal Data about the potential and real users, their lifestyle, web usage and search histories, social media linkages, personal interests and hobbies, age and income groups, neighbourhoods and travel profiles. Specialised algorithms are then deployed to mimic personal search for every user. Such enormous amounts of data, its collection and use, have grave implications on the future of journalism and social ethics. The data thus collected can not only be used for profiling users and selling them to advertisers but it can also be deployed for political campaigning, targeting propaganda and for manipulating the media and the elections.

8.5.1 Data regulation and Protection as a Right of Citizens

We know from the above discussion that it is common for tech companies to do regular and real-time data snooping and the use of algorithms to customise information for individual users. It is precisely for this reason that data protection laws are required. Data protection in India is poor mainly because it is relatively new area for citizens as well as governments. However, it is important for data journalists to understand the need and ramifications of data protection. It has also become important in the light of widespread complaints of cybercrimes and data

(and identity) theft. Personal data of the citizens is often stolen from the banks and insurance companies. It is often seen that even government controlled data is sometimes leaked or stolen. One of the biggest problems is the absence of an effective framework of data protection.

In Aug. 2017 the Supreme Court of India made privacy a fundamental right through a landmark order. The Apex Court warned against the threat to privacy from both state and non-state actors, asking the government to put in place "a robust regime for data protection." The existing data protection laws (i.e. the Information Technology Act 2000 and its rules framed in 2011) cover only corporates, not government entities, and only sensitive personal data, such as medical history, biometric information etc.[14] Ultimately, there is lack of recourse for citizens – breach of personal information is not ground enough for seeking redress, except in case of financial loss.[15] The absence of a regulator continues to be a key weakness of the system. Justice BN Srikrishna committee set up in 2017 to suggest an effective data protection regime proposed Personal Data Protection Bill 2018. It is hoped that the data protection will be an important part of citizens' rights in the future.

8.6 HDRS AND DATA ABOUT HUMAN DEVELOPMENT

Human Development Reports (HDRs) are published by UNDP annually to evaluates the countries on the basis of their performance on the Human Development Indicators (HDI): Life Expectancy, Knowledge and Income (per capita income) of the population. A composite index is formed from ranking the countries under the three indicators and a combined list of final ranking of the countries is published with HDR for which data is collected round the year. As per the Global Human Development Report 2016: Human Development for Everyone, India is ranked at 130 among the 188 countries. HDR comes up with open-ended, sustainable solutions to increase human development through contemporary means. The HDRs were later published in the form of National HDRs, State HDRs and even district HDRs (for selected districts) which collate enormous amount of data about human indicators.[16] On the basis of these, a country can conduct self-assessment of their performance on HDI and frame constructive policies to enforce improvement. State-level HDRs have also become a fundamental part of planning and political discourse for the majority of the Indian states in last 20 years.

8.6.1 Annual Status of Education Reports (ASER)

Annual Status of Education Report (ASER) is an autonomous research and assessment unit, established by Pratham, a civil society organisation that engages in estimating children's schooling and learning status at district, state and national levels. ASER survey reaches almost every rural district in the country and covers more than 15,000 households and 650,000 children each year[17]. ASER is an annual survey that aims to provide reliable estimates of children's enrolment and basic learning levels for each district and state in India. ASER has been conducted every year since 2005 in all rural districts of India. It is the largest citizen-led survey in India[18]. ASER is trying to generate actionable evidence and on the basis of that, solutions to deal with the situation can be identified. This evidence is derived on a variety of data collated from surveys, assessments, evaluations and other research activities in education and other social sectors. For instance, the

2017 survey examines what the students do, what they know, and what they want besides the school enrolment & learning abilities.

8.6.2 India Exclusion Reports

India Exclusion Report (IXR) is a multi-disciplinary report published by Centre for Equity Studies. It stands as a platform for social scientists, policy researchers, scholars, etc. to promote inclusion in accessing rights and government services. The IXR report often discovers inconsistencies between the needs of people and the allocation of resources by the government. The IXR, 2016 is the third edition of the report that highlights the systemic exclusion of certain groups, despite several declarations by the government. The report also looks at the impact of economic policies from the perspective of the vulnerable sections such as the Scheduled Castes, Scheduled Tribes, the minorities, women and persons with Disabilities, who were historically underprivileged. The report includes other vulnerable groups such as the urban street children and the manual scavengers. The report also discusses India's position in the Global Hunger Index and the state of malnutrition in the country.

8.7 TIMES-SERIES DATA

Time Series data, as the name suggests, is a sequence of data arranged or listed in order of their time. It is usually collected over a very long time and presented in the form of lists, tables or graphs which can be measured and subjected to meaningful analysis of trends, patterns and extrapolation of data to make predictions about the future. For instance, if data is available of, say, 100 years of rainfall or weather, one can find out cycles of extreme patterns and even predict about the future. Time series is used in economics, weather, rainfall, census, earthquakes etc. Many data journalists with mathematics, statistics or econometrics backgrounds can make sense of data more easily but a refresher course can familiarise almost all students in understanding and analysing long-term data. This allows students to develop data analysis skills as part of a framework of critical thinking.

8.7.1 Following exercises may prepare students to look at the time-series data critically:

1. Watch the TED video link below of the famous data scientist, physician and statistician Hans Rosling and write 500 words on the co-relations between population and prosperity around the world along with supporting data.
2. https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen?language=en
3. Go to the website of the Census of India and spot trends and patterns about the co-relations between population and prosperity in India on the above lines
4. Visit the website of the National Sample Survey Organisation (NSSO) which conducts large-scale sample surveys in diverse fields and write a note on the range of subjects covered.
5. See the latest Annual Report of the Reserve Bank of India (on the RBI website) and prepare a note on the state of the economy and what it says about the data on investment climate and growth in the past year

6. Go through the latest Economic Survey and list five most critical areas of growth in the Indian economy in the last year

7. List five most important types of data collected by the National Family Health Survey (NFHS) and write a short note on its methods of data collection

8. Go through the website of the Bureau or Police Research and Development (BPRD) and the National Crime Records Bureau (NCRB) and make a list of five most important types of data collected about crime and prisons in India.

9. Go through the exit polls conducted by different agencies and media organisations during the last Assembly elections in India and compare them for accuracy.

Check Your Progress 2

Note: 1) Use the space given below for your answer
2) Compare your answer with those given at the end of this Unit.

1. Explain how does data regulation can help in protecting the citizens' rights?

.....

.....

.....

.....

.....

8.8 SIMPLIFYING THE CHALLENGES OF DATA JOURNALISM

One of the lessons of data journalism is that it is skill oriented work which can be learnt by following a method. It is never easy to deal with numbers when they come in huge quantities or in the form of time-series. However, it is no rocket science either; even the most accomplished researchers and data scientists ensure data integrity, and check and recheck data before coming to conclusions. What must be always kept in mind is that data is not something unusual but it is a part of everyday life. It is impossible to understand government policies without an understanding of data. It is also difficult to nail the truth about environmental issues like polluting factories of climate change or medical data like cancer research and spread of diseases without deploying long-term data. Fortunately, there are enough web-based resources available to modern journalism students to hone their skills. It is a huge relief that once basic proficiency is acquired in making sense of data, the process can be very enlightening and empowering. No politician or bureaucrat can deceive or delude a data journalist by bombarding her with data.

8.8.1 Some Important Indian Websites/ Outlets Useful For Data Journalism

a) India Spend http://www.indiaspend.com/

b) Inclusive Media for Change http://www.im4change.org/

c) Peoples' Archive of Rural India (PARI) https://ruralindiaonline.org/

d) South Asia Terrorism Portal http://www.satp.org/

e) Centre for Science and Environment https://www.cseindia.org/

f) E-Social Sciences (online scholarly resource)

http://www.esocialsciences.org/Home/NewIndex.aspx

8.8.2 Important International web Resources on Data Journalism

Resources for Learning and Doing Data Journalism (Press Institute of America database) https://www.americanpressinstitute.org/publications/reports/strategy-studies/data-resources/

"Knowing the Numbers," (Harvard's Journalist's Resource project) http://isoj.org/research/knowing-the-numbers-assessing-attitudes-among-journalists-and-educators-about-using-and-interpreting-data-statistics-and-research/

Video: "Solve Every Statistics Problem with One Weird Trick," NICAR 2016, Jonathan Stray https://www.youtube.com/watch?v=3TIJ6KVlmF0

The Upshot (The New York Times) https://www.nytimes.com/section/upshot

Stanford Computational Journalism Lab http://cjlab.stanford.edu/

Data Is Beautiful, a community on Reddit https://www.reddit.com/r/dataisbeautiful/

Data is Plural. Sign up for http://tinyletter.com/data-is-plural. All datasets can be found in an updated master spreadsheet.

Jonathan Stray, The Curious Journalist's Guide to Data, 2016. https://towcenter.org/research/the-curious-journalists-guide-to-data/

Brant Houston, Computer-Assisted Reporting: A Practical Guide, 2014.

David Herzog, Data Literacy: A User's Guide, 2016. https://study.sagepub.com/herzog

ProPublica https://www.propublica.org/

Video: "The Newest Muckrakers: Investigative Reporting in the Age of Data Science," Sarah Cohen, C+J Symposium Stanford, 2016

http://journalism.stanford.edu/cj2016/#livestream

8.9 LET US SUM UP

This Unit has covered certain core skills and a broader understanding of data, an analysis of its sociology and politics, and a range of benefits arising out of its presentation and interpretation. A precondition to practicing data journalism is to understand that the art and science of data journalism requires proficiency in specific skills, right from simple articulation of numbers to an understanding of spreadsheet programmes to its visualisation and interpretation for meaningful analyses. The same has been simplified and conveyed to you through this unit.