Statistical Ethics for Data Science

Jefferson Evans, Thomas Sperry
Center for Social Impact, Women in Business Impact Lab
Utah Valley University

ABSTRACT

Misrepresenting information and understanding is a common issue with statistics. However, with the rapid advancement of AI, concerns surrounding deepfakes, fake news, and the misuse of statistics are more widespread than ever. Increasing statistical literacy and promoting public awareness is essential in combating these problems.
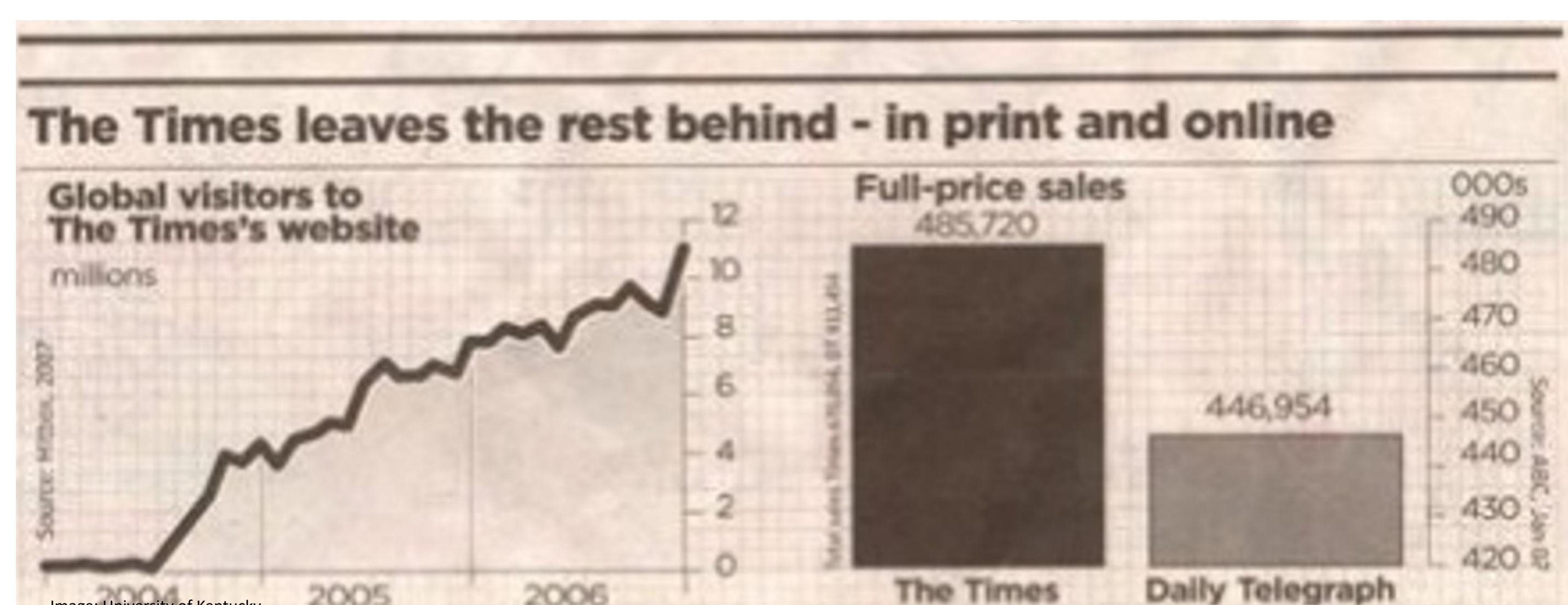
INTRODUCTION

Throughout our research this semester it quickly became evident how easily statistics could be misrepresented. Raising awareness on the topic is the quickest way to combat the issue. Data science is a rapidly growing field that focuses on real world applications of statistics. Understanding the issues surround data science will greatly mitigate future problems.

MISUSE OF STATISTICS

Statistics whether simple or complex carry a lot of power. Descriptive statistics such as the mean and median are often used to describe a large amount of data simply. However, even using such well-known statistical tools incorrect conclusions can be drawn. The severity of misuse only increases with more sophisticated tools are incorrectly implemented.

Businesses have been knowingly or unknowingly misleading others with misused statistics for decades. Pictured below is a case where "The Times" newspaper used a specific scale for their chart to make it seem as though they were drastically outperforming their competitors. Taking a closer look, you can see how that the information is misleading.

WHAT IS DATA SCIENCE?

Data science is a field that utilizes statistics and computer science. Data science uses computers to apply advanced statistical methods on large data sets. The insights gathered from the analysis increase optimization, allow for more accurate predictions to be made, and machine learning to occur. Many industries rely heavily on observations from this analysis.

As the world becomes increasingly interconnected gathering and processing large quantities of data becomes more important. As the industry grows, so must the focus on accurate analysis.

Data science is a tool that can be used to help protect marginalized groups and make all industries more equitable. However, the same information can be used to increase inequality and ostracize communities as well. As the value of data increases so will the ethical concerns over privacy and data protection.

ISSUES

- Increased automation of fake news
- Confirmation bias
- Increased data privacy concerns
- Misapplication of statistical methods
- Lack of critical thinking
- Deepfakes

With the rapid progress of data science and AI tools there are many concerns and issues. It will be easier than ever to create and widely spread fake news. Users of social media and other apps will unknowingly have more sophisticated tools decided what they see and when leading to confirmation bias. Personal data will be worth more than ever and will need to be protected.

AI tools carry a wealth of knowledge that is very useful for learning and applying math, statistics, and science. However, without a proper understanding of what is being used erroneous conclusions may be drawn due to a lack of understanding of the mathematical or scientific tools applied.

On the left, Saturday Night Live star Kate McKinnon impersonates Elizabeth Warren during a skit, and on the right, face swap deepfake technology has been used to superimpose Warren's face onto that of McKinnon.
Credit: Stephen McNally/UC Berkeley

RECOMMENDATIONS

- Develop technologies to detect misused statistics
- Increase the number of courses teaching media literacy
- Criminalizing malicious use
- Protecting intellectual property
- Regulating how AI is developed and implemented
- Disclosing when AI or deep fakes are involved
- Creating liability for companies that promote or misuse AI

New technologies bring rapid growth that can benefit everyone. The use of such technologies must be carefully implemented in order to avoid the negative impacts that can come as well.

REFERENCES

- https://www.universityofcalifornia.edu/news/new-technology-helps-media-detect-deepfakes
- Wheelan, C. (2014). Naked statistics: Stripping the dread from the data. W. W. Norton & Company.

ACKNOWLEDGEMENTS