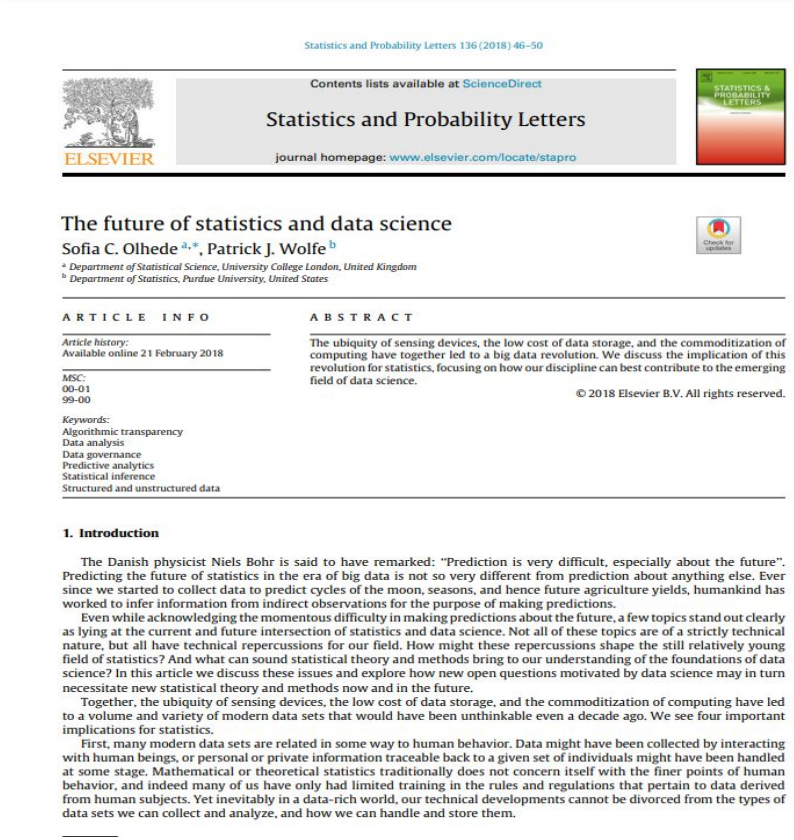
# The future of statistics and data science (2018)



Statistics and Probability Letters 136 (2018) 46–50

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

## The future of statistics and data science

Sofia C. Olhede [a,*], Patrick J. Wolfe [b]

[a] Department of Statistical Science, University College London, United Kingdom
[b] Department of Statistics, Purdue University, United States

ARTICLE INFO

*Article history:*
Available online 21 February 2018

*MSC:*
00-01
99-00

*Keywords:*
Algorithmic transparency
Data analysis
Data governance
Predictive analytics
Statistical inference
Structured and unstructured data

ABSTRACT

The ubiquity of sensing devices, the low cost of data storage, and the commoditization of computing have together led to a big data revolution. We discuss the implication of this revolution for statistics, focusing on how our discipline can best contribute to the emerging field of data science.



### 1. Introduction

The Danish physicist Niels Bohr is said to have remarked: "Prediction is very difficult, especially about the future". Predicting the future of statistics in the era of big data is not so very different from prediction about anything else. Ever since we started to collect data to predict cycles of the moon, seasons, and hence future agriculture yields, humankind has worked to infer information from indirect observations for the purpose of making predictions.

Even while acknowledging the momentous difficulty in making predictions about the future, a few topics stand out clearly as lying at the current and future intersection of statistics and data science. Not all of these topics are of a strictly technical nature, but all have technical repercussions for our field. How might these repercussions shape the still relatively young field of statistics? And what can sound statistical theory and methods bring to our understanding of the foundations of data science? In this article we discuss these issues and explore how new open questions motivated by data science may in turn necessitate new statistical theory and methods now and in the future.

Together, the ubiquity of sensing devices, the low cost of data storage, and the commoditization of computing have led to a volume and variety of modern data sets that would have been unthinkable even a decade ago. We see four important implications for statistics.

First, many modern data sets are related in some way to human behavior. Data might have been collected by interacting with human beings, or personal or private information traceable back to a given set of individuals might have been handled at some stage. Mathematical or theoretical statistics traditionally does not concern itself with the finer points of human behavior, and indeed many of us have only had limited training in the rules and regulations that pertain to data derived from human subjects. Yet inevitably in a data-rich world, our technical developments cannot be divorced from the types of data sets we can collect and analyze, and how we can handle and store them.


Mohammad NASR

mo.t.nasr21@gmail.com

## Sofia C.Olhede

Professor of Statistics, Department of Statistical Science, University College London, United Kingdom

- High-dimensional Statistics
- Bias

## Patrick J.Wolfe

Professor of Statistics, Department of Statistics, Purdue University, United States

- Cited by 4359
- Signal Processing
- Machine Learning

# Layout

1. Introduction

2. Missing the data science boat?

3. Data governance

4. Regulation and algorithmic transparency

5. Structured and unstructured data

6. Bias, incompleteness, and heterogeneity

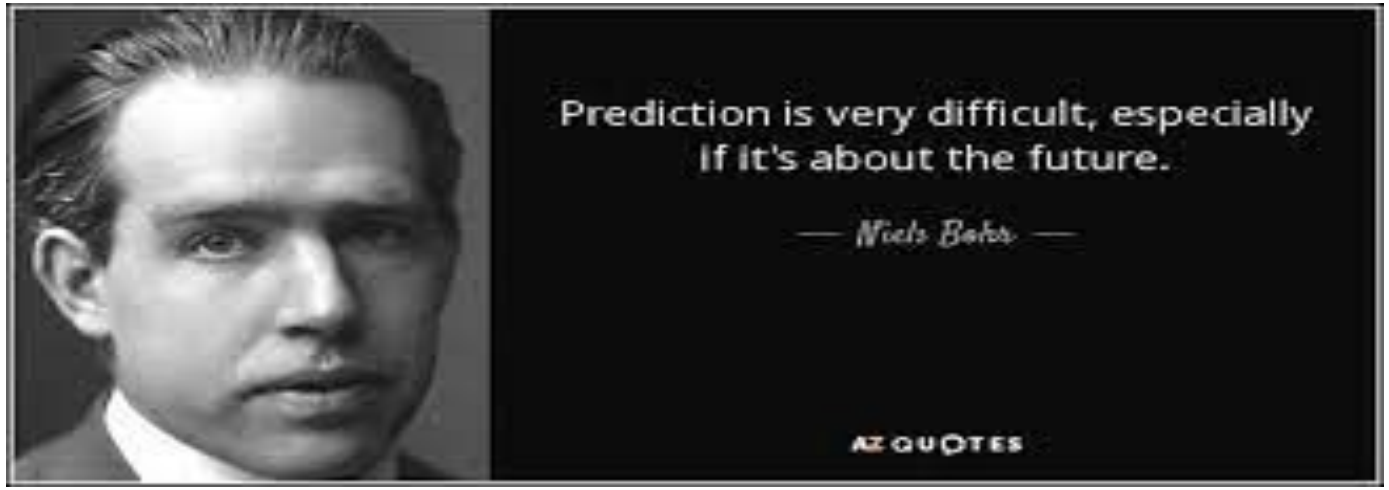
7. Discussion

# Keywords

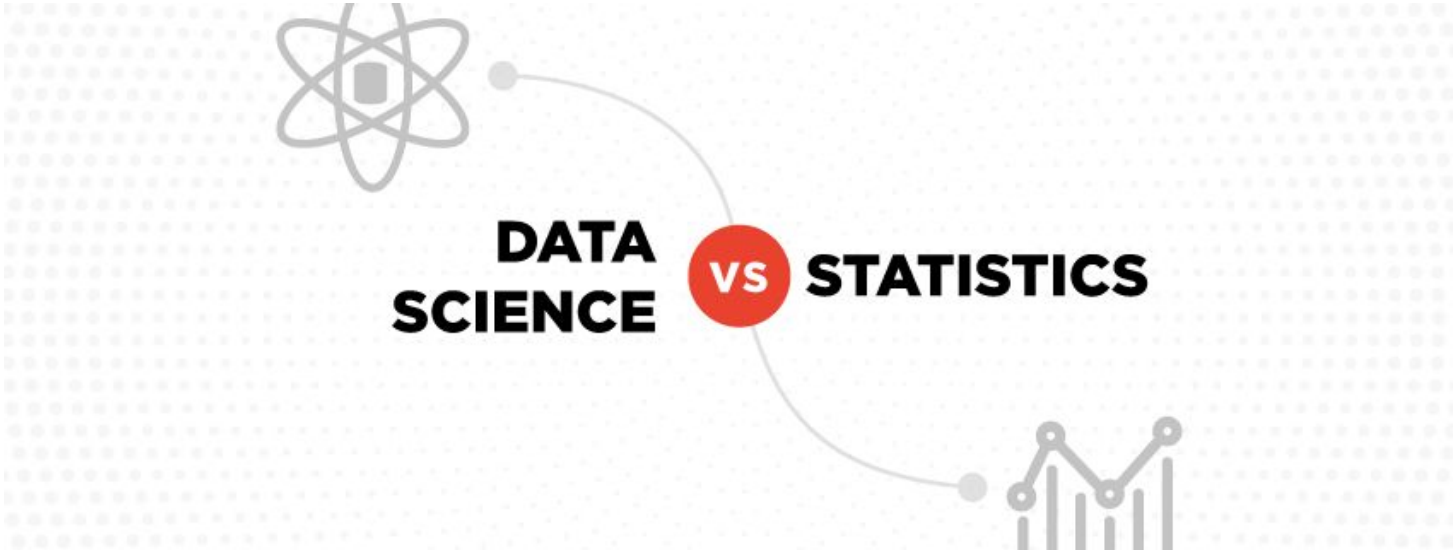
- ***Algorithmic Transparency:***
  - ***Definition:*** Decision-making processes of an algorithm are clear. Transparent algorithms allow users to comprehend how they arrive at specific outcomes.
- ***Data Analysis:***
  - ***Definition:*** Inspecting, cleaning, transforming, and modeling data.
- ***Data Governance:***
  - ***Definition:*** Policies that ensure high data quality, integrity, security, and compliance throughout an organization.
- ***Predictive Analytics:***
  - ***Definition:*** Identifying patterns in historical data and make predictions about future events or trends.
- ***Statistical Inference:***
  - ***Definition:*** Drawing conclusions or making predictions about a population based on a sample of data
- ***Structured and Unstructured Data:***
  - ***Structured Data:*** Databases with a well-defined schema.
  - ***Unstructured Data:*** Information that lacks a predefined data model or structure.

# 1. Introduction

- The future of statistics in the era of big data appears in danger at first glance.
- The intersection of statistics and data science.
- Human behaviour in theory not taken into account
- Data regulation is required
- Growing complexity of Data
- IID Distribution not met



# 2. Missing the data science boat?



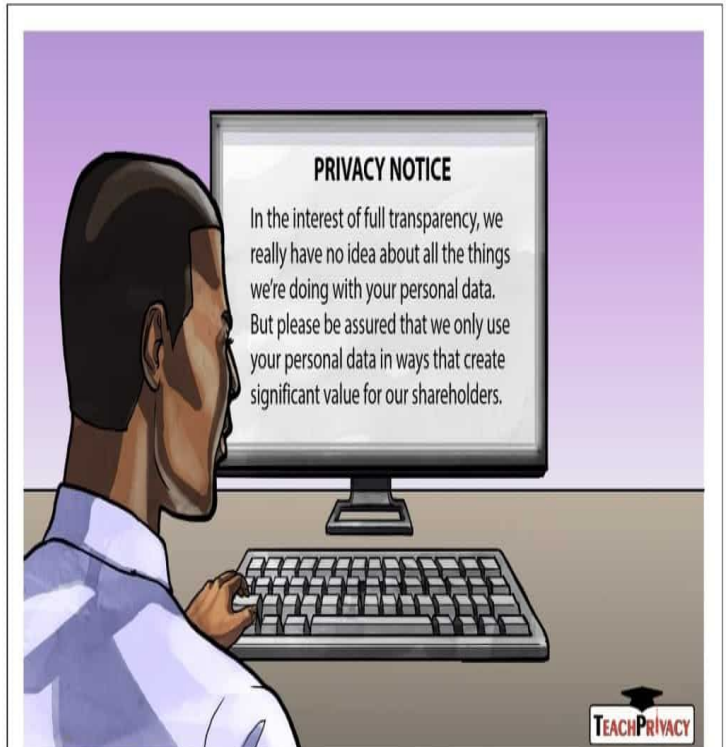

***-> Thirty-eighth Conference on Stochastic Processes and their Applications in 2015*** debate: "This house believes that the mathematical scientists will miss the data science boat"

- The argument for statistics missing the data science boat centered on statisticians favoring theoretical challenges.
- The counterargument suggesting that the boat would sink without it.
- The debate ended in a draw, reflecting the valid points on both sides.

# 3. Data governance

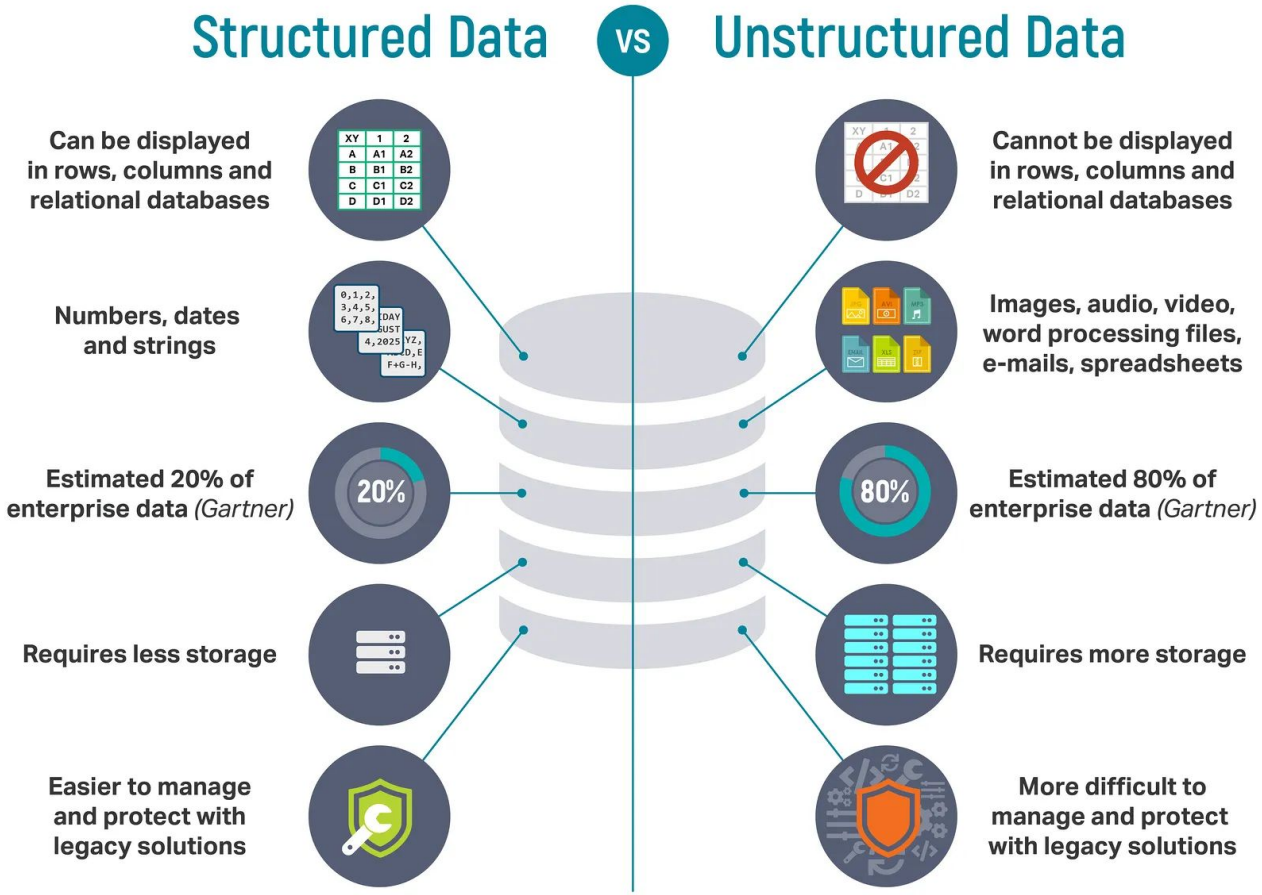
- Immense societal opportunities but also significant threats.
- Crucial to avoid misuse of data and build public trust.
- Designing fail-safe anonymization schemes and analyzing anonymized data.
- (IEEE): Institute of Electrical and Electronics Engineers
- British Academy and UK Royal Society

# 4. Regulation and algorithmic transparency




- Data Collection and Governance Concerns on the rise.
- Regulation of Algorithms is required.
- Call for Algorithmic Transparency.
- Challenges in Achieving Transparency.
- Understanding Prediction Processes.
- Issues with Predictive Algorithm Complexity.
- Importance of Interpretability Tools.
- Balancing Predictive Accuracy and Transparency.

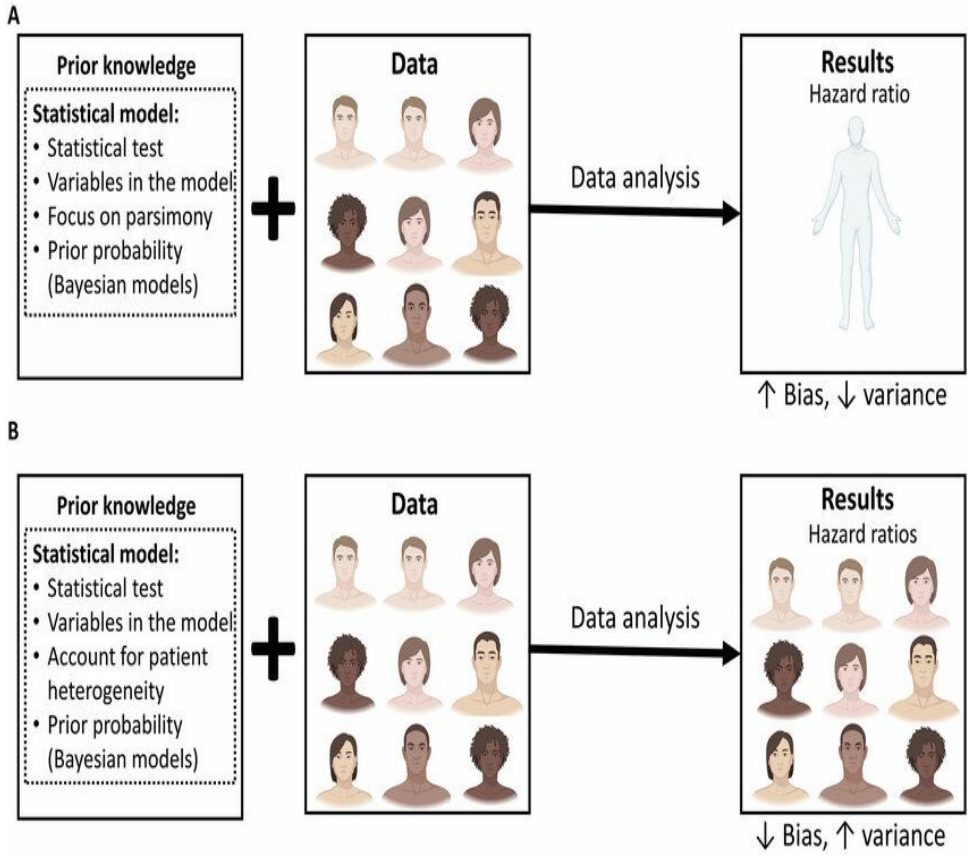
# 5. Structured and unstructured data




- Increasing complexity of data (Big Data) and the need for new modern statistical tools..
- Statistical Challenges in Network Analysis.
- Bridging the Gap Between Data and Models.
- Computational Challenges in Network Analysis.
- Research Goals in Network Analysis, from proving results to establishing limits.

# 6. Bias, incompleteness, and heterogeneity

- Missing and biased observations pose critical challenges in data science.
- Access to "all" data doesn't eliminate the need for statistical models or sampling methodologies.
- Underdeveloped Statistical Theory, working with limited “found sampling data.
- Correlation between Observation and Inference.
- Analysis of populations with high heterogeneity is a significant challenge.
- Effectively modeling heterogeneity and understanding the sampling replication properties of complex random objects remain open problems.
- Tailored Modeling for Inferential Foundation



# 7. Discussion

**The current focus in data science has been on predictive "black box" tools rather than classical modeling and the need for statistics and data science to work hand in hand with the rise of big Data and algorithmic complexity while respecting the social factor**