50 Years of Data Science

Author: David Donoho(2017)

Justin Li

11/04/2024

Abstract

In 50 Years of Data Science, David Donoho critically examines the development of data science as a distinct field, tracing its evolution from statistical foundations and highlighting the influence of pioneers like John Tukey, who advocated for a broader focus on data analysis over mathematical statistics. The paper explores the growing divergence between traditional statistics and data science, noting how the latter emphasizes real-world data handling, computational proficiency, and scalable methodologies for "big data." Donoho addresses how modern data science programs increasingly prioritize skills in data management, machine learning, and prediction over traditional inference, reflecting industry needs. His critique is that contemporary data science may neglect foundational scientific rigor by focusing on immediate commercial applications rather than developing a comprehensive academic discipline. In my presentation, I will discuss Donoho's proposed framework for a holistic "Greater Data Science," encompassing six areas from data gathering to scientific validation of data practices, as well as his vision for a field that balances empirical methodologies with foundational insights, setting a path for data science's future in both academia and industry.

Introduction

David Donoho’s paper, "50 Years of Data Science," reflects on the history, current state, and potential future of data science.

- Purpose: Differentiate data science from statistics.

- Critiques current commercial focus of data science.

Proposes a vision for data science with scientific rigor.

About the Author

David L. Donoho
- Position: Professor of Statistics at Stanford University.
- Notable Contributions:
  - Pioneer in statistical theory, signal processing, and data science.
  - Developed methods for compressed sensing, wavelets, and noise reduction.
  - Advocate for open science and reproducible research.
- Recognition:
  - Member of the National Academy of Sciences and American Academy of Arts and Sciences.
  - Recipient of prestigious awards, including the Shaw Prize in Mathematical Sciences.

Background and Historical Context

Data science's roots trace back to John Tukey’s 1962 paper, "The Future of Data Analysis." Tukey envisioned data analysis as a field distinct from mathematical statistics. Advances in computing and the explosion of data have since shaped the field into what we now call "data science."

What is Data Science?

Donoho defines data science as a field that goes beyond traditional statistics, involving data collection, management, processing, and prediction. Data science aims to make data useful, prioritizing practical applications over purely theoretical insights.

Definition:

- - Data science is the field focused on making data useful.
- - Involves data collection, cleaning, processing, visualization, and prediction.
- - More application-driven than traditional statistics.

The "Data Science Moment"

The past decade has seen the rise of data science programs at universities like UC Berkeley and MIT, driven by industry demand for big data skills. Data science is now recognized as a distinct field, with specialized degree programs and career paths.

Data Science vs. Traditional Statistics

Key differences between data science and traditional statistics include a focus on prediction rather than inference, an emphasis on large datasets ("big data"), and the need for extensive computing skills.

Key Differences:

- Data Science: Focuses on prediction and working with large, complex datasets.
- Traditional Statistics: Focuses on inference, understanding relationships, and smaller datasets.
- Data science requires significant computing skills, beyond classical statistics.

The Six Divisions of Greater Data Science

Donoho proposes "Greater Data Science" (GDS), dividing it into six core areas:

1. Data Gathering, Preparation, and Exploration
2. Data Representation and Transformation
3. Computing with Data
4. Data Modeling
5. Data Visualization and Presentation
6. Science about Data Science

Division 1: Data Gathering, Preparation, and Exploration

This division includes collecting data, cleaning and preparing it for analysis, and performing exploratory data analysis (EDA) to gain initial insights.

Involves:

- Collecting raw data from various sources.
- Cleaning data to address inconsistencies and missing values.
- Using Exploratory Data Analysis (EDA) to understand data patterns and issues.

Division 2: Data Representation and Transformation

Data scientists often need to transform data into usable formats. This includes converting text data to numerical form, applying mathematical transformations, and using data structures like SQL and noSQL databases.

Key tasks:

- Reshape and convert data to suitable formats (e.g., text to numerical data).
- Use tools like SQL or noSQL databases for structured storage.
- Apply transformations to enhance data features (e.g., log transformations, Fourier transforms).

Division 3: Computing with Data

This division covers programming languages like Python and R, as well as computational tools for managing and analyzing large datasets on cloud or distributed systems.

Skills required:

- Proficiency in programming languages
- Understanding of cloud computing and distributed systems for handling large datasets.
- Familiarity with data processing frameworks like Hadoop or Spark.

Division 4: Data Modeling

Data modeling includes both predictive modeling (focused on accuracy) and inferential modeling (focused on understanding). Machine learning often falls under this category in data science.

Types of modeling:

- Predictive Modeling: Prioritizes accuracy in forecasting outcomes.
- Inferential Modeling: Focuses on understanding data relationships and causality.
- Machine learning algorithms are commonly used for predictive tasks.

Division 5: Data Visualization and Presentation

Data visualization is essential for communicating results. Good visualizations make complex data insights accessible and support decision-making.

Importance:

- Visualization makes complex data understandable.
- Essential for communicating findings to non-expert audiences.
- Tools like Tableau, Matplotlib, and D3.js are commonly used.

Division 6: Science about Data Science

This area involves studying data science practices themselves, evaluating their effectiveness, and refining methodologies to improve the field.

Focus:

- Study of data science practices and methodologies.
- Aim to evaluate and improve the effectiveness of data science techniques.
- This meta-approach ensures data science itself is based on evidence.

The Future of Data Science

Donoho advocates for a data science that goes beyond commercial applications, focusing on scientific rigor and intellectual growth. He envisions a field that addresses complex, interdisciplinary challenges.

Conclusion

In summary, data science has grown into a field distinct from traditional statistics, yet it must continue to evolve. Donoho’s vision for a rigorous, evidence-based data science emphasizes the importance of intellectual growth and scientific integrity.

References:

- - Donoho, D. L. (2017). *50 Years of Data Science*. Journal of Computational and Graphical Statistics, 26(4), 745-766. DOI: 10.1080/10618600.2017.1384734
- - Tukey, J. W. (1962). *The Future of Data Analysis*. Annals of Mathematical Statistics, 33(1), 1-67.
- - Shaw Prize Foundation. *Laureate Profile: David L. Donoho*. Retrieved from https://www.shawprize.org/laureates/mathematical-sciences
- - National Academy of Sciences. *Member Directory: David L. Donoho*. Retrieved from https://www.nasonline.org/member-directory/members/20006252.html
- - Additional readings on data science: Cleveland, W. S. (2001). *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*. International Statistical Review, 69(1), 21-26.