50 Years of Data Science

Overview on the paper by David Donoho: Department of Statistics, Stanford University

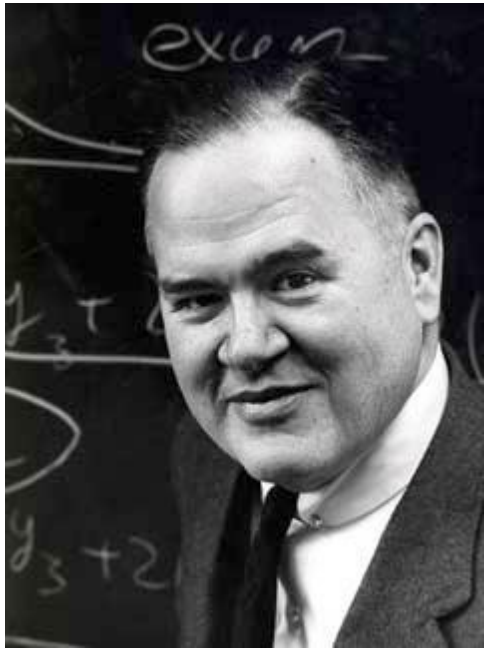
Ahmed Abderraouf Menaa
05/10/2023

ahmedmenaapro@gmail.com

Today's Data Science Moment

50 years ago. **John Tukey**, an American Mathematician and Statistician, called for a reformation of academic statistics. He identified the existence of a science focused on learning from data, which he called "data analysis".

- In September 2015, the university of Michigan announced a 100-million-dollar data science initiative.
- Campus wide initiatives at NYU, Columbia, MIT.
- New master's degree programs in data science in NYU, Stanford.



Data Science or Statistics?

Data Scientist: is a professional who uses scientific methods to liberate and create meaning from raw data.

Statistics: means the practice or science of collecting and analysing numerical data in large quantities

- "Aren't we Data science?" ~ President Marie Davidian in AmStat News
- "A grand debate: is data science just a 'rebranding' of statistics" ~ Martin Goodson, co-organizer of the Royal Statistical Society on May 19, 2015.
- "Data Science without statistics is possible, even desirable" ~ Vincent Granville, at the Data Science Central Blog10
- "Statistics is the least important part of data science" ~ Andrew Gelman, Columbia University

Data Science or Statistics?

Big Data meme: Statisticians rejected the term big data due to two facts:

- History: the term Statistics was coined 200 years ago when countries tried to collect data about its residents. This is today's equivalent of 'Big Data'.
- Science: Statisticians have been researching how to deal with large datasets for years, they focused on understanding their behavior for both large observations or measurements.

Skills meme: statisticians also rejected the accusation that they don't have the necessary skills for data analysis

- Wayback, statisticians could fit databases on a single processor and use simple functions like minimum, maximum, and mean easily
- While other complex tasks or large-scale optimization were manageable using elegant mathematics.

The Future of Data Science, 1962

"For a long time, I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt." ~ John Tukey

- He foreshadows the emergence of data science, with his ideas relevant to arrival of data science.
- He argues that the revolution of data science is beyond scaling up, but we have to make it a field where we learn from the data.

50 Years after FoDA

Impact of John Tukey's "The Future of Data Analysis":

- Influence of Tukey's ideas on individual statisticians, such as P.J. Huber.

Slow Adoption of Tukey's Vision:

- Slow acceptance within academic statistics departments.
- Divide between statisticians focusing on mathematical statistics and those embracing data analysis.

Current State of Data Science Programs:

- Data science faculty at institutions like UC Berkeley have diverse backgrounds.
- Data science perceived as a distinct field that extends beyond statistics.

Emphasis on Open Science:

- Scientific publication should enable reproducibility.
- Computational results and data analyses should be fully described and accessible.
- Emphasis on transparency and reproducibility in data science research.

Breiman's Two Cultures

Generative Modelling Culture:

- Focuses on developing stochastic models fitting the data.
- Aims to make inferences about the underlying data-generating mechanism.
- Assumes the existence of a true model generating the data.

Predictive Modelling Culture:

- Emphasizes prediction over inference.
- Focuses on accurate predictive algorithms.
- Prioritizes prediction accuracy on various datasets.

The predictive Culture's Secret Sauce

Common Task Framework (CTF):

- Includes publicly available training dataset, competitors, and scoring referee.
- Competitors develop prediction rules based on training data.
- Referee evaluates submitted rules using a testing dataset.
- Promotes competition and small incremental improvements..

The Full Scope of Data Science

Essential Skills in Data Science Education:

1. Data Gathering: data preparation, understanding messy data with cleaning.
2. Data Representation: understanding of various data formats and transformations
3. Computing with Data: knowledge of programming languages and cluster computing

The Full Scope of Data Science

Essential Skills in Data Science Education:

4. Data Modelling: deep understanding of both generative and predictive modelling cultures

5. Data Visualization: visualizations for insights using advanced techniques

The Full Scope of Data Science

Essential Skills in Data Science Education:

A data science program checks all points and provides:

- Broader curriculum
- Hands on experience with data
- Research: developing R tools, and other practical contributions that improve effectiveness or efficiency

The next 50 Years of Data Science

Open Science takeover:

- Research reproducibility will be most important to verify findings and improve quality
- Empirical testing of Data analysis algorithms
- Performance of algorithms across a range of real-world databases will determine the best algorithms
- This evidence will outweigh mathematical derivations and proof

Conclusion

- data science represents an enlargement of academic statistics and machine learning. The field of data science has its roots in data analysis and modeling, and it has expanded to encompass various aspects of data processing and technology.

- the core motivation for the expansion into data science is intellectual. While there may be industrial demand for data science skills, the primary driving force behind the field is scientific rather than industrial

- the scope and impact of data science will continue to expand significantly in the coming decades. With the availability of vast amounts of scientific data.