

“A Century of Data Science”

The evolution of data science has completely transformed how the world views data and the insights drawn from it. From just numbers or text, the data has now become the biggest asset to achieve our desired tasks in our daily businesses. Either it's finance, medicine, engineering or computing, in short no matters which field it is the data has become our need. But, the fact is that data itself is meaningless and unusable, unless it is properly processed and cleaned. For this very task, from the earliest days when data was recorded on paper to today's era of well-structured digital databases, statistics has remained our backbone. With its immense range of data preprocessing and manipulation techniques, statistics has always provided the tools needed to turn raw information into meaningful insight. Over time, this statistical foundation expanded, matured, and ultimately evolved into the modern field we now know as “Data Science.”

In this blog, I will cover complete journey of data cleaning and processing since 1925, highlighting all the important inventions, dates and scientists who played vital role in shaping data science into the discipline we know today.

MileStones of Data Science:

1. Phase 1: The Statistical Era (1925–1945)

These two foundational decades are known for their inventions on both statistical and data science techniques that are the basis of our today's advance data science and statistical techniques. Some important inventions and dates dedicated to these decades are:

- a. **The Fisherian Revolution:** In 1925, British statistician **Sir Ronald Fisher** publishes “Statistical Methods for Research Workers”. He introduces the concepts of the "p-value", ANOVA testing, correlation and regression techniques, that gives scientists a mathematical way to determine how much the given data is useful to them.
- b. **The Universal Machine for Data Manipulation:** In 1936, a British mathematician **Alan Turing** publishes a research paper “On Computable Numbers”. He conceptualizes a machine that can perform any calculation if right instructions are given, the theoretical ancestor of every computer today.
- c. **The First Artificial Neuron:** In 1943, **Warren McCulloch** and **Walter Pitts** publish a paper modeling a biological neuron as a digital logic gate. This is the Big Bang of Neural Networks.

2. **Phase 2: The Dawn of Digitalization & Databases (1946–1980)**

This era is considered as one of the important turning points of the history. As the business related data were started storing and processing electronically making it more efficient and accurate. Some important inventions and dates dedicated to these decades are:

- a. **ENIAC**: In 1946, first general-purpose electronic computer was built by John Presper Eckert and John Mauchly. It could calculate artillery trajectories in seconds, proving that machines could process data thousands of times faster than humans.
- b. **The Turing's Test**: In 1950, **Alan Turing** proposes a test: if a human cannot tell they are talking to a machine, the machine is "intelligent." This set the ultimate goal for AI.
- c. **The Hypothesis of John Tukey**: In 1962, **John Tukey** publishes a landmark paper arguing that statistics needs to evolve into a new science focused on exploring data, not just proving theorems. He effectively predicts modern Data Science.
- d. **Relational Databases (SQL)**: In 1970, **Edgar Codd** at IBM publishes the model for Relational Databases. This allowed data to be stored in tables with rows and columns and accessed them using a structured language. It remains the backbone of banking and business database systems.

3. **Phase 3: The Information Age & Big Data (1981–2010)**

In this period, Internet connected the world and an immense amount of unstructured and inconsistent data (text, images etc.) that cannot be handled by the traditional databases. Some important inventions and dates dedicated to these decades are:

- a. **The World Wide Web**: In 1989, **Tim Berners-Lee** invents the World Wide Web, to fix the problem of sharing information, storing data under a same roof using hyperlinked documents that anyone could access through a browser.
- b. **Python**: In 1991, **Guido van Rossum** releases Python. Decades later, it would become the primary language of data science due to its simplicity and powerful libraries (like Numpy, Pandas, Matplotlib etc.).
- c. **Hadoop**: In 2006, **Doug Cutting** releases Hadoop. It allowed companies to process "Big Data" by splitting it across thousands of cheap servers. This enabled Facebook, Yahoo, and Amazon to scale their business models.

4. *Phase 4: Era of AI & Generative Intelligence (2011–2025)*

From very recent Deep Learning allows machines to teach themselves.

The focus shifts from analyzing the past to generating the future.

- a. **The AlexNet Breakthrough:** In 2012, **Geoffrey Hinton** and his team knocked down the competition ImageNet (computers classifying images into categories) using a Deep Convolutional Neural Network. This proved that "Deep Learning" was the future of computer vision.
- b. **The Transformer Architecture:** In 2017, Google researchers publish "Attention Is All You Need". They invent the Transformer model, which allows computers to understand given context in any language. This is the architecture behind ChatGPT.
- c. **ChatGPT (Generative AI):** In 2022, OpenAI releases ChatGPT, making advanced AI available to the general public for the first time. The focus shifts from predictions (of ideas) to the generation (of answers on demand).
- d. **Small Language Models (SLMs) and Edge AI:** In 2025, Scientists has advised the world not to rely only on massive cloud based AI models, but also on developing Small Language Models (SLMs) and Edge AI, designed to run directly on laptops, phones, and local devices.

Conclusion

Over a century, the evolution of data science has **completely transformed** human capabilities. A discipline rooted in simple statistics and manual data gathering has expanded into a **global intelligence system** powered by massive datasets, distributed computing, and advanced artificial intelligence. Each achievement, whether the invention of relational databases, the rise of machine learning, or the emergence of generative models, has pushed humanity closer to understanding, predicting, and shaping the world with more precisely than ever before. Today, data science is no longer limited to field related to data but also a foundation of modern advancements, influencing every aspect of life from finance to creativity. As we dive deep into data science, we find that the journey of a whole century was built not on the basis of data and its manipulation, but on the human curiosity and innovation giving actual meaning to data.

References

1. Alan Turing, 1936, "The Birth of Computing"
2. Ronald A. Fisher, 1925, "Statistical Methods for Research Workers"

3. The First Neural Network, 1943, "Warren McCulloch & Walter Pitts"
4. John Tukey, 1962, "The Birth of Data Analysis"
5. Tim Berners Lee ,1989 , Information Management: A Proposal
6. Claude Shannon ,1948 , "A Mathematical Theory of Communication"
7. Sundas Khalid ,Senior Data Scientist at Google
8. Amir Sepehri, Senior Data Scientist at Meta
9. Tom Davenport, Senior Author of Data Analytics