Thinking Hardware in Data Science

High Performance Data Analytics (HPDA): Trends and Persepctive

Oluwaseyi Ayanda[1] Kunle Oladosu[1] Taiwo Kehinde[1] Ruth Aliche[1]

[1] School of Engineering, Computing and Mathematics, University of Plymouth, UK

Introduction

High performance data analytics (HPDA) is the use of high-performance computing (HPC) to analyse large data sets for patterns and insights **(HPE, no date)**. HPC systems provide the technology and computational power required to process and analyse big data workloads **(Pattanshetti, 2020)**. HPDA unites HPC with big data analytics **(Pattanshetti, 2020)**.

They do this by using multi-processors that are networked together for parallel and distributed computing **(IBM, 2024)**. The characteristic nature of big data (velocity, volume, variety) has influenced Convergence and synergy between HPC and data science **(IBM, 2024)**. HPDA aims to combine the capabilities of HPC with big data analytics to address the challenges posed by big data **(Pattanshetti, 2020)**.

Why HPDA?

- Some analytics workloads do better with HPC rather than standard compute infrastructure;
- some analytics workloads can only be run with HPC;
- sensitive timeframe for analysis (real-time);
- Down sampling dataset leads to suboptimal result **(NVIDIA, 2024)**.

Big Data Challenges & Characteristics

VOLUME: THE SIZE OF THE DATA.

VELOCITY: THE SPEED AT WHICH THE DATA IS GENERATED.

VARIETY: THE DIFFERENT TYPE OF DATA.

VERACITY: THE TRUSTWORTHINESS OF THE DATA IN TERMS OF ACCURACY.

VALUE: JUST HAVING BIG DATA IS OF NO USE UNLESS WE CAN TURN IT INTO VALUE.

Figure 1: Big Data Challenges & Characteristics. Source (Julian M. Kunkel. HPDA2022)

Exponential growth in the worlds data

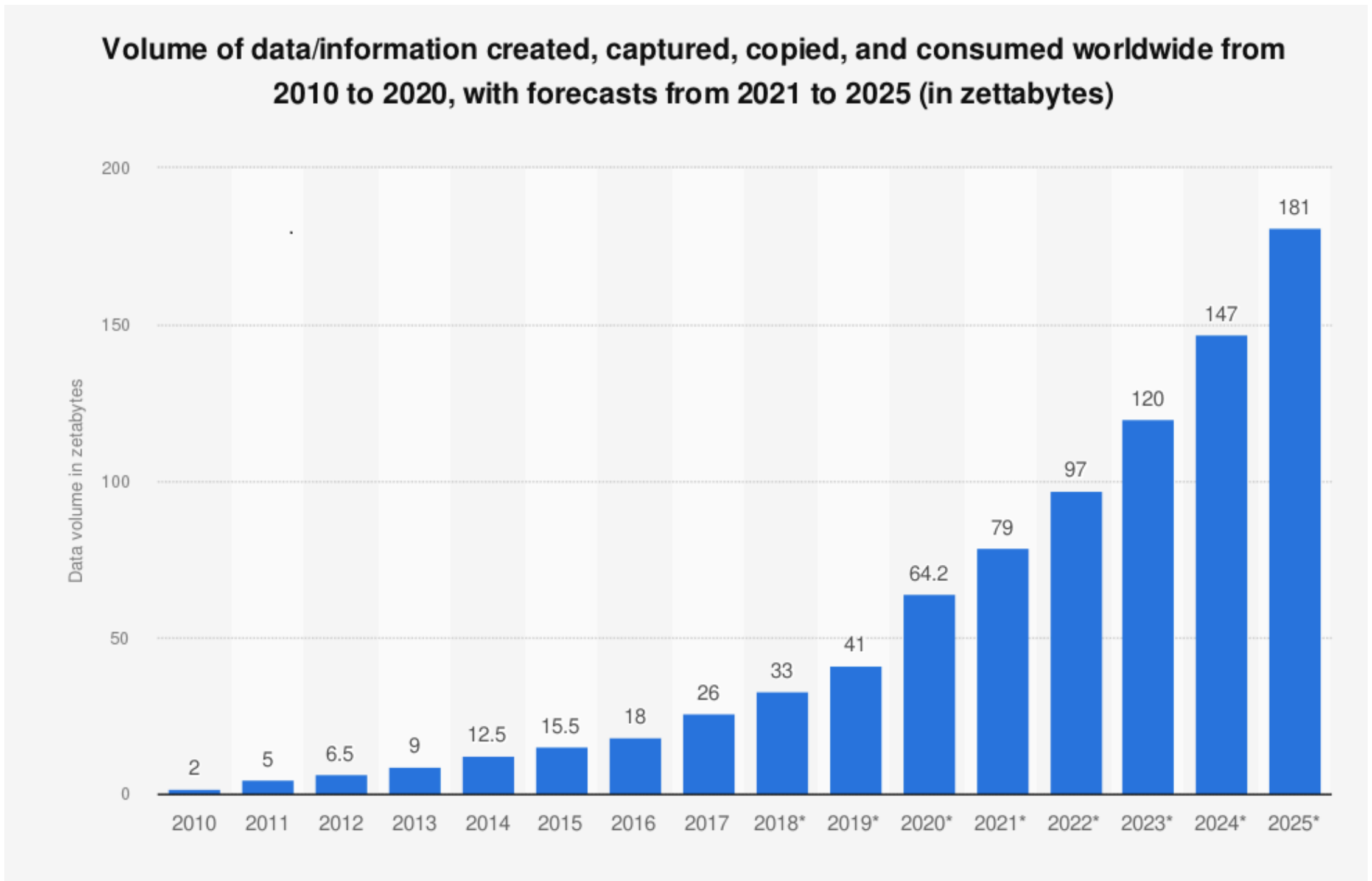


Figure 2: Rapid proliferation and unprecedented growth in the volume and diversity of the worlds data. Credits (https://www.statista.com/statistics/871513/worldwide-data-created/)

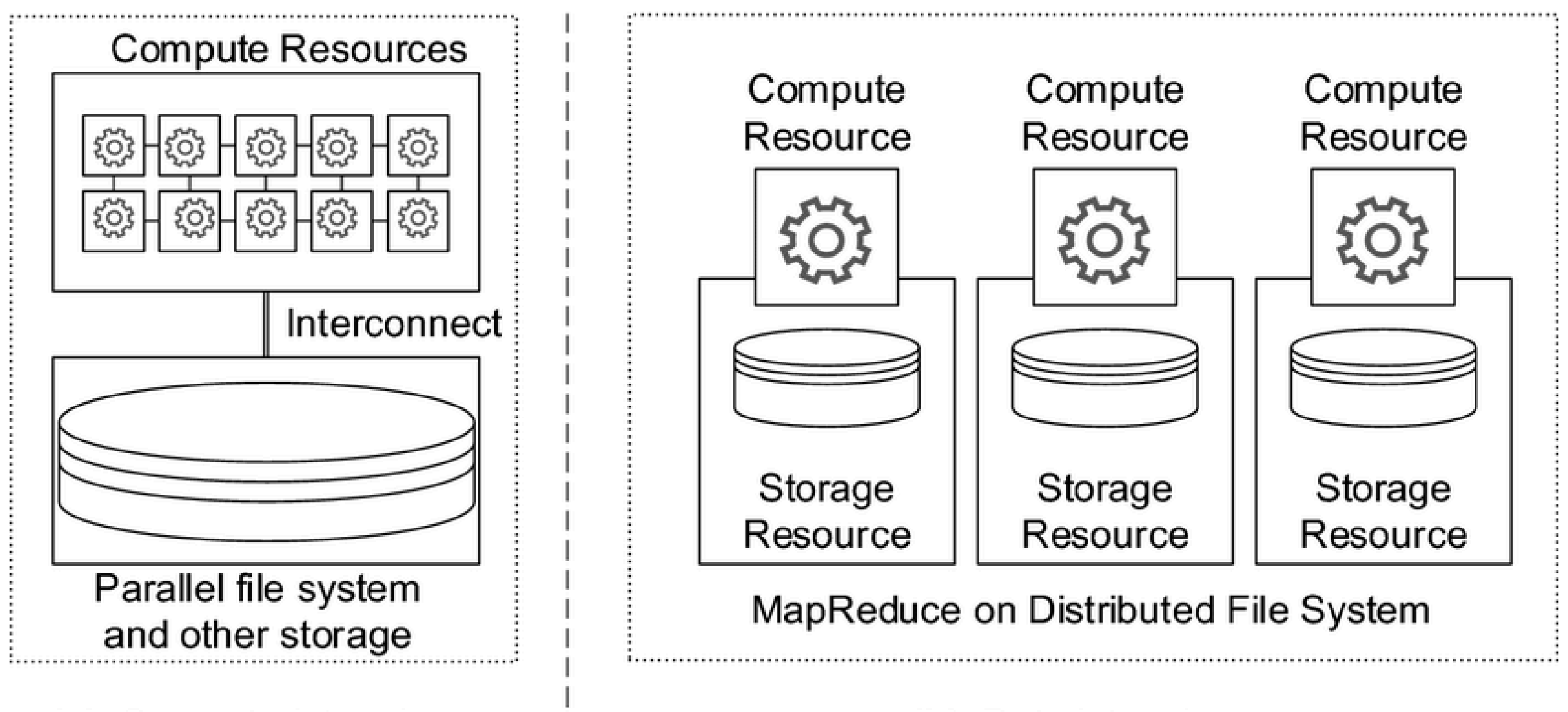
HPC vs Big Data Analytics System

Table 1: High Performance Computing vs Big Data Analysis ***(Caino-Lores et al., 2019, Pattanshetti, 2020)***

High Performance Computing	Big Data Analytics
Optimised for complex computational processes on smaller data sets.	Optimised for simpler computations on large volumes of data.
Use Specialized hardware.	Use commodity hardware.
Focus on high-speed data processing.	Focus on handling large volumes of data from various sources.
How can we cope with increasing datasets?	How can we run analytics faster?

Computing Architecture for HPDA

- `Data Intensive Process (shared memory)`: Comprise several compute nodes, connected together with a high-performance network, along with login nodes (login servers within the HPC Cluster) and an external file system **(Jackson *et al.*, 2019)**.
- `Compute Intensive Process (distributed memory)`: Comprise several compute nodes with non-shared private memory, connected together with a high-performance network, along with login nodes and/or external file system **(Pathak, Pandey and Rautaray, 2020)**.



(**a**). Compute-intensive processing paradigm

(**b**). Data-intensive processing paradigm

Data science techniques for HPDA

- Deep learning (DL): a subset of machine learning (ML) that employs artificial neural networks with multiple layers to learn data representations. Common use cases include include image recognition, natural language processing, and speech recognition.

HPDA Use Cases

- **Digital twins in manufacturing**: Simulation of realistic problems where the simulation model is expected to provide its results synchronously with the real world **(Ejarque *et al.*, 2022)**.
- **Financial services**: Financial modeling, portfolio optimization **(IBM, 2024)**
- **Cyber security**:
  - Fraud detection: Analysing an event against millions of records to find a potentially fraudulent transaction in real time.
  - Surveillance, monitoring of large-scale infrastructures **(HPE, no date)**.
- **Ecommerce**: Design of recommendation engines that proritises personalization **(Newman, 2023; LENOVO, no date)**.
- **Weather Forecasting and Climate Modelling**: processing vast amounts of historical meteorological data and millions of daily changes in climate-related data points.
- **Healthcare, genomics and life sciences**: Drug discovery and design **(IBM, 2024)**.
- **Energy sector**: Design of Smart energy grid.
- **Design of intelligent systems**: Autonomous vehicles.

Analysis

Breakthrough in scientific computing is radically reshaping the landscape of data science

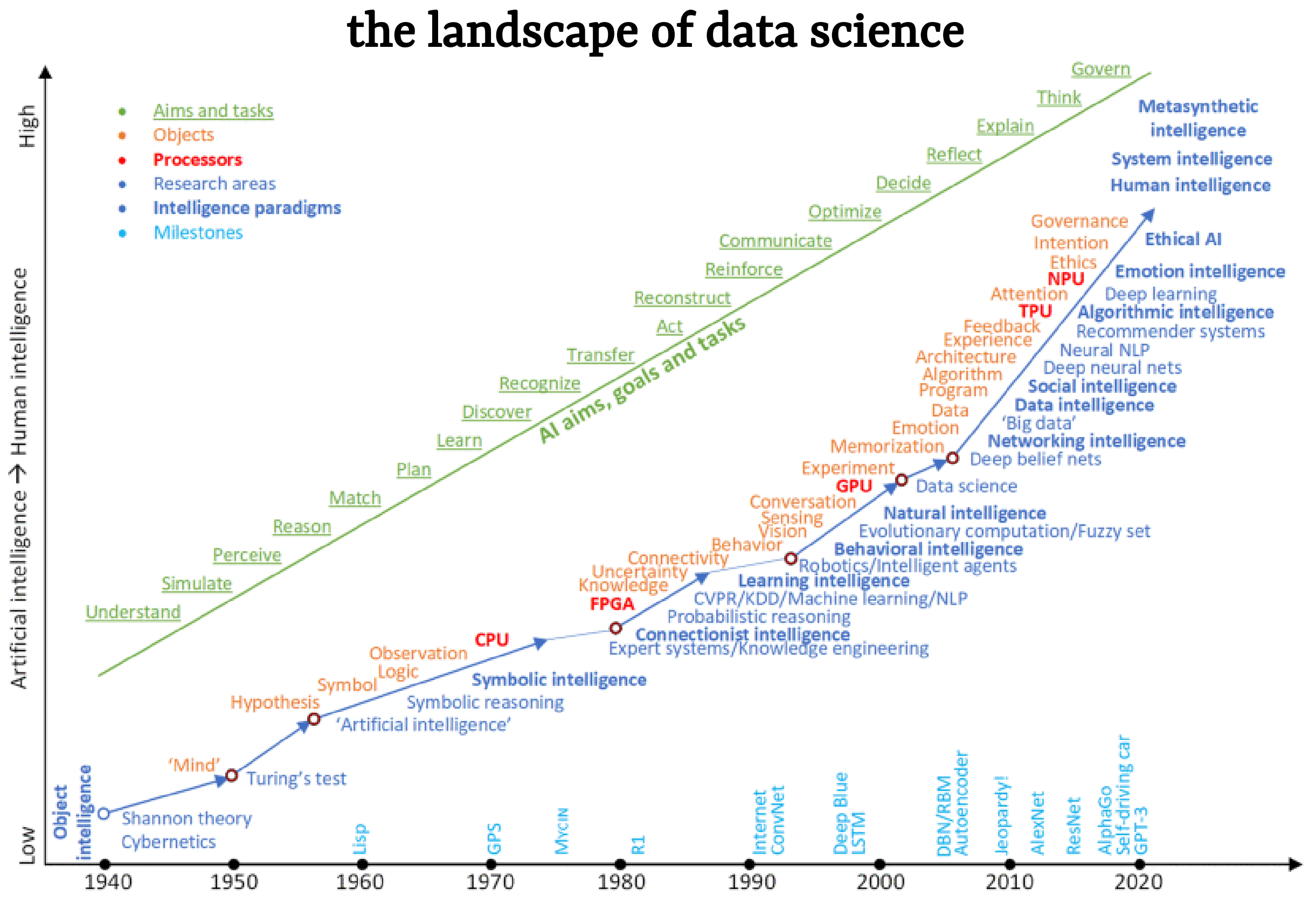


Figure 3: Breakthrough in computing systems and data science knowledge areas ***(Cao, 2022)***.

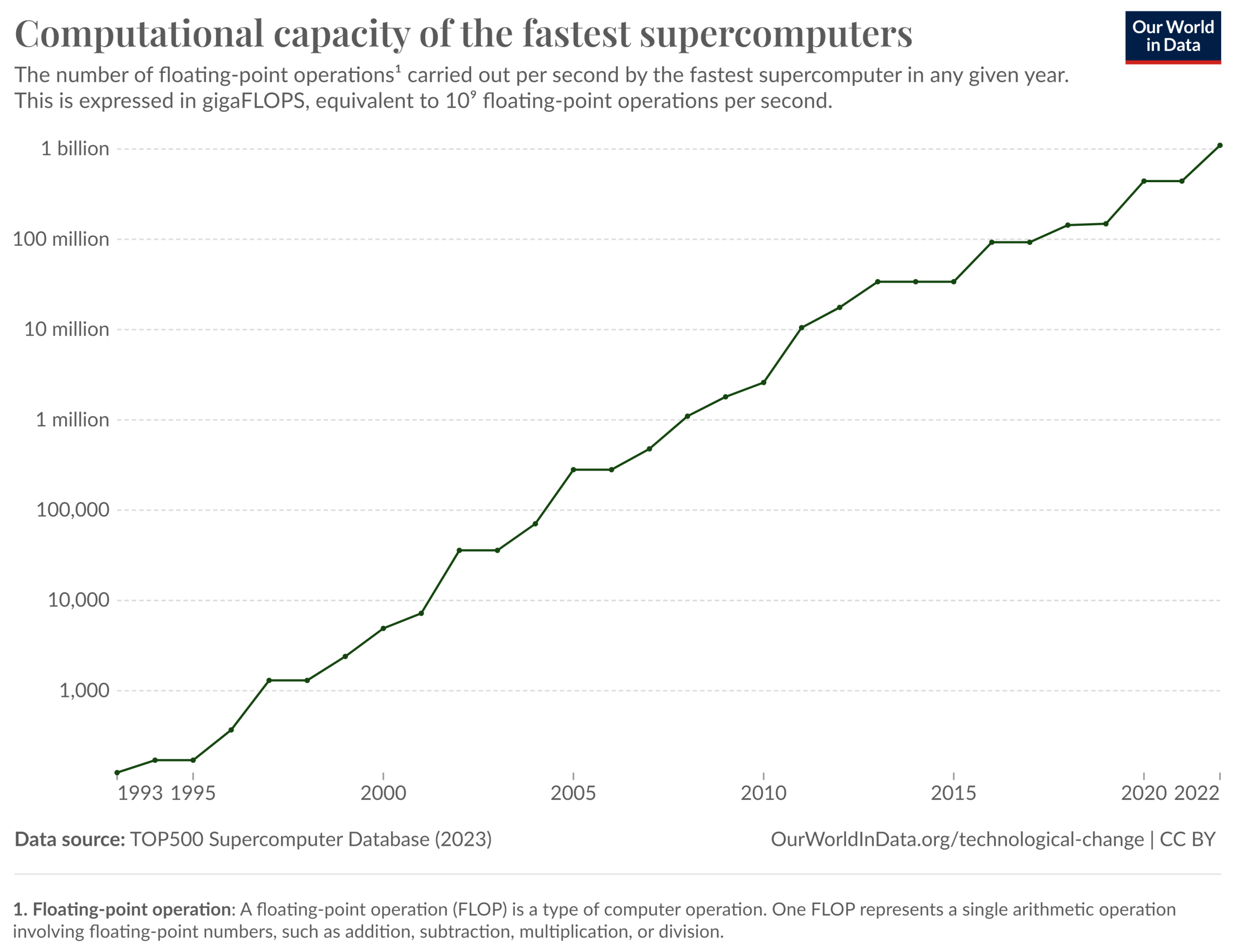


Figure 4: Computational capacity of the world fastest supercomputers. Credits(https://ourworldindata.org/grapher/supercomputer-power-flops)

Future directions

- HPDA will see increased applications in data science. This will be driven by the exponential increase in data, the lower cost of advanced computing systems, and the quest to uncover insights from big data analysis **(Asch *et al.*, 2018)**.
- Development of personalized products and services, e.g., personalized health care.

Conclusion

- Every organization should think critically about the hardware it chooses to support its data analytics workload now and in the future. The continuous growth of data and the development of new analytical models demand this **(Cao, 2022)**.
- In this era of the fourth industrial revolution and Exascale computing, HPDA is the key for businesses and organizations to solve their ever-growing computational and data-intensive tasks.

References

Asch, M. *et al.* (2018) 'Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry', *The International Journal of High Performance Computing Applications*, 32(4), pp. 435–479.
Cao, L. (2022) 'A new age of AI: Features and futures', *IEEE Intelligent Systems*, 37(1), pp. 25–37.
Ejarque, J. *et al.* (2022) 'Enabling dynamic and intelligent workflows for HPC, data analytics, and AI convergence', *Future generation computer systems*, 134, pp. 414–429.
HPE (no date) 'What is high performance data analytics?' https://www.hpe.com/uk/en/what-is/high-performance-data-analytics.html/.
IBM (2024) 'What is high-performance computing (HPC)'. https://www.ibm.com/topics/hpc.
Jackson, A. *et al.* (2019) 'An architecture for high performance computing and data systems using byte-addressable persistent memory', in. Springer, pp. 258–274.
Kunkel, J.M. (2023) 'High performance data analytics lecture'. https://hps.vi4io.org/_media/teaching/autumn_term_2022/hpda22-13.pdf.
LENOVO (no date) 'What is HPC?: Types of high-performance computing models'. https://www.lenovo.com/us/en/faqs/servers/what-is-high-performance-computing/?orgRef=https%253A%252F%252Fwww.google.com%252F.
Newman, D. (2023) 'The future of personalization: What you need to know', *Forbes*. Forbes Magazine. Available at: https://www.forbes.com/sites/danielnewman/2023/05/07/the-future-of-personalization-what-you-need-to-know/.
NVIDIA (2024) 'End-to-end accelerated analytics with nvidia'. https://www.nvidia.com/en-us/deep-learning-ai/solutions/data-analytics/.
Pathak, A.R., Pandey, M. and Rautaray, S.S. (2020) 'Approaches of enhancing interoperations among high performance computing and big data analytics via augmentation', *Cluster Computing*, 23(2), pp. 953–988.
Pattanshetti, M.K. (2020) 'Design of high-performance computing system for big data analytics', *JOURNAL OF ALGEBRAIC STATISTICS*, 11(1), pp. 115–121.
Thorne, W.B. (2019) *Posterdown: An r package built to generate reproducible conference posters for the academic and professional world where powerpoint and pages just won't cut it*. Available at: https://github.com/brentthorne/posterdown.

UNIVERSITY OF PLYMOUTH