The Role of Data: Collection, Storage, Manipulation, and Movement

Anthony J. Coots

National University

TIM-8102: Principles of Computer Science

Dr. ********

April 27th, 2025

The Role of Data: Collection, Storage, Manipulation, and Movement

Data is omnipresent in computer science; it is key to every digital interaction and computational process. When a user presses a power button on a personal computer, data is collected, stored as an electrical signal, manipulated by the corresponding software and hardware layers, and moved across the entire system to continue operations. It is a mundane action to the average user, yet a replica of how modern computing systems work with and act on data. It is the age of big data and artificial intelligence, and data volume, velocity, and variety continue to grow exponentially. Data collection, storage, manipulation, transformation, and movement are necessary and strategic processes for data-driven decision-making, data analytics, and "intelligent" behavior among machines. This paper intends to explore data in its current nature and relation to computer science.

Data, Information, and Metadata

Data

Data, the plural datum, is raw and unprocessed facts representing a state. It is a factual representation of things without any context or interpretation. For example, "yellow" in different contexts could be a color, a password, and so forth, as interpreted; on its own, it simply denotes a property with no meaning. Data only gains significance when it is processed or organized meaningfully.

Information

When data is given context or interpreted, it becomes information. Only then does "yellow" go from being a data point to information, such as a car's color, a house, or a password.

Information prescribes understanding from raw facts. It is used to gain knowledge and make decisions.

Metadata

Metadata is universally described as “data about data.” It provides context and structure to raw data, making it easier to locate, understand, and manage. Metadata is fundamental among programmers, database administrators, and other computer science professionals.

National University (personal communication, April 2025), metadata can be categorized into three forms: descriptive, structural, and administrative.

- Descriptive: “… to identify data resources based on characteristics assigned to the data.”
- Structural: “… to establish parts that are attributed to the whole set of data in a logical group.”
- Administrative: “… to manage resources through the creation of rules, constraints or restrictions, and guides or instructions for the data.”

Examples in Computer Science

Li et al. (2025) explore how metadata can be used with deep learning, a form of artificial intelligence in computer science, for data-driven decision-making during relevance evaluation. The researchers developed a model called the Integrative Use of Metadata for Data Sense-Making, which aimed to identify how users integrate systematic and heuristic metadata cues through two cognitive patterns, within-category and across-category integration. The team tracked users’ eye movements and used an interpretable deep learning model to validate patterns during relevance assessments of datasets. Ultimately, they found that integrating metadata, when used by intelligent systems, assisted users when evaluating the relevance and utility of data, cementing the importance of data, information, and metadata.

Almadi et al. (2025) utilize machine learning to monitor fetal health by developing multi-class classification models using cardiotocogram data. Their study uses over 2,000 cardiotocogram exam records classified into three health states: normal, suspect, or pathological. They implemented several models, such as Logistic Regression, Random Forest, and Neural Networks, to predict fetal health based on information such as fetal heart rate, uterine contractions, and histogram accelerations and decelerations. Their research ultimately showed high accuracy for critical cases, demonstrating that when data is used appropriately, it can create insight for healthcare professionals to improve maternal outcomes.

Purpose of Stages in the Data Life Cycle

Creation

The data life cycle begins with data creation. Without the generation or gathering of data, no subsequent steps such as storage, use, sharing, archiving, or destruction occur. In other words, the data life cycle does not exist without data creation. By extension, computer-driven processes or digital systems cannot perform any tasks unless the system is specifically designed to do just that, nothing. With this in mind, data creation often involves surveys, study participation, and the translation into structured, semi-structured, or unstructured data formats.

Storage

After data is created, it must be stored in a manner appropriate to its purpose. On computers, data can be stored in a variety of ways. If stored via electrical signals, it may reside in volatile or non-volatile memory, depending on whether power is required. Non-volatile memory, for example, retains data without power and often in common file types such as PDFs or CSVs. This retained data can be organized in structured, semi-structured, or unstructured formats, depending on the use case. Azad et al. (2020) organize data storage mechanisms in the Internet of Things with SQL databases for structured data, NoSQL databases for semi-structured data, and graph databases for unstructured and highly connected data, each serving different storage needs.

Use

One of the most potent components of data therein lies in how it is used to generate information to support decision-making. For example, many conceptual ideas can now be substantiated with data-driven models when utilized in machine learning. For instance, Qushem et al. (2024) demonstrate using data in predictive analytics to identify at-risk students in computer science programs. The study developed and evaluated multiple machine learning models using student course performance data across four academic years, where the best model even achieved an accuracy of 96%. This data, which uses course grades as input, is processed through machine learning algorithms, and the output predictive insight illustrates how data can be used to create actionable information, in just one example. When shared with academic faculty or students, this could influence a student’s educational career trajectory for the greater good.

Sharing

Data sharing is an important function in the life cycle, as it allows information to be distributed and utilized for decision-making and collaboration. However, it also demands careful attention to privacy, ethics, and regulatory compliance. For example, individuals working in academic institutions must adhere to the Family Educational Rights and Privacy Act (FERPA). At the same time, those in the healthcare sector must comply with the Health Insurance Portability and Accountability Act (HIPAA). As the value of data continues to increase, so does the risk of violating federal law, cementing that secure sharing methodology is ever so important.

Furthermore, data sharing occurs at many levels. At a high level, it may be passing data via physical media such as external drives. On a lower level, it may be transmitted wirelessly from machine to machine using technology such as RFID. Specific to the Internet of Things, Azad et al. (2020) show how modern data-sharing mechanisms involve interconnected devices communicating across systems. Thus, frameworks must be able to handle sharing high volumes of data.

Archiving or Destruction

These final stages in the data life cycle, archiving and destruction, are often considered together because of their contracting yet interconnecting nature. As a working database administrator, one is frequently faced with a decision once data has been accepted as input, processed into a relational database management system, and the necessary output has been produced.

Archiving involves preserving data for future reference, regulatory compliance, or historical analysis. Depending on the operational needs and system design, this can occur on a schedule, by the minute, hour, day, and so on. Archived data is typically stored in compressed formats or transferred to long-term storage solutions to maintain accessibility while reducing resource demands, such as drives on servers in a network.

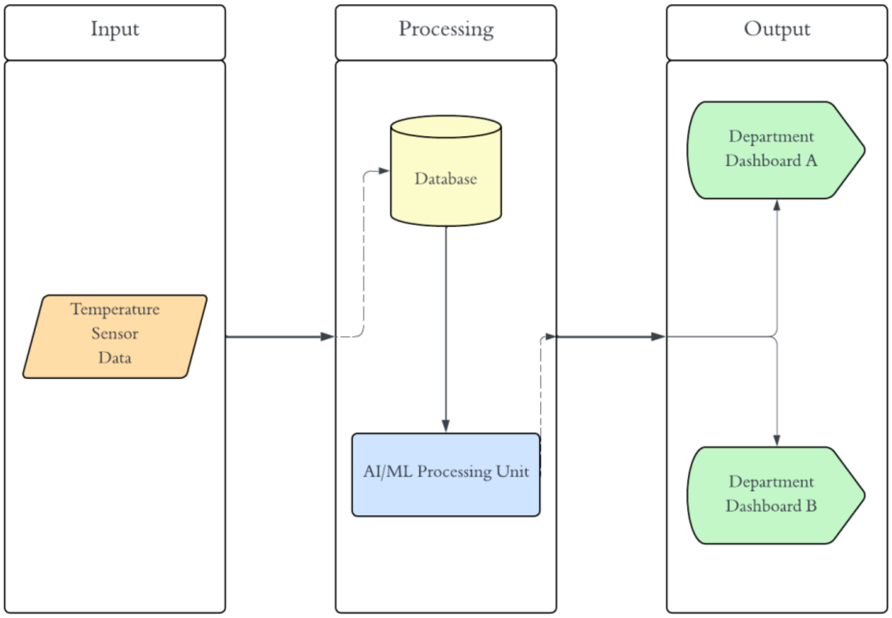
In contrast, destruction is the secure deletion of data that is no longer needed. Whether data is archived or destroyed depends on retention policies and expectations, which are usually defined at the beginning of the data life cycle. As any extenuating law applies, choosing between the two paths depends on system efficiency, data integrity, and general compliance.

Hypothetical Input, Processing, and Output

Data Flow Diagram

A hypothetical data flow begins by taking raw data as input, processing it into a format suitable for storage within a database, and then analyzing it using an artificial intelligence or machine learning processing unit or program. For example, this system could notify an Information Technology department if a server room becomes hot or alert a facilities department if a particular area on campus exceeds a temperature threshold. These notifications are the output, which started with sensor input and were appropriately processed. See Figure 1 for a visual representation of this flow.

Figure 1. Data Flow Diagram for Temperature Monitoring as Hypothetical Input, Processing, Output.



Conclusion

Data is everywhere in all areas of computer science, at its simplest levels it is the hardware interactions given an electrical signal, and at more complex levels it is used for machine learning systems. Real-life examples, such as the prediction of student success and fetal health classification using machine learning, demonstrate how data can be utilized to participate in systems for beneficial outcomes. As data-driven technology continues to evolve, the ability to understand and apply data strategically remains important for research and operations in computer science.

References

Almadi, M., Alotaibi, F., Almudawah, R., Ali, A., Nasser, Y., & Nasser, N. (2025). Data-Driven Machine Learning Models for Enhanced Fetal Health Classification and Monitoring. *2025 8th International Conference on Data Science and Machine Learning Applications (CDMA), Data Science and Machine Learning Applications (CDMA), 2025 8th International Conference on, CDMA*, 189–192. https://doi.org/10.1109/CDMA61895.2025.00038

Azad, P., Navimipour, N. J., Rahmani, A. M., & others. (2020). The role of structured and unstructured data managing mechanisms in the Internet of Things. *Cluster Computing*, *23*(2), 1185–1198. https://doi.org/10.1007/s10586-019-02986-2

Li, Q., Wang, P., Liu, C., Li, X., & Hou, J. (2025). Integration patterns in the use of metadata for data sense-making during relevance evaluation: An interpretable deep learning-based prediction. Journal of the Association for Information Science & Technology, 76(3), 621–641. https://doi.org/10.1002/asi.24961

National University. (2025). *Data as information* [Course handout]. NCUOne.

Qushem, U. B., Oyelere, S. S., Akçapınar, G., Kaliisa, R., & Laakso, M.-J. (2024). Unleashing the power of predictive analytics to identify at-risk students in computer science. *Technology, Knowledge and Learning: Learning Mathematics, Science and the Arts in the Context of Digital Technologies*, *29*(3), 1385–1400. https://doi.org/10.1007/s10758-023-09674-6