Concept 6: Big Data

Since the beginning of the Internet, humans have contributed an immense volume of data—billions upon billions of data points. Each point carries a piece of information, and when aggregated, this information forms what is commonly called **Big Data**. While "Big Data" may refer broadly to all digital information, in practice it often denotes the data collected in a specific context—for instance, an online retailer's detailed record of customer purchasing patterns, or a loyalty program's tracking of consumer spending and rewards.

Remarkably, it is estimated that **90% of today's Internet data has been generated during the past two years**. We now create as much data every 48 hours as was accumulated from the dawn of humanity until the year 2000. The growth is exponential: the Internet already stores approximately **5 zettabytes** of data, and experts predict this will increase tenfold by 2020.

Big Data encompasses all forms of digital information—from text and databases to images, sound recordings, and real-time sensor inputs. These datasets are so vast and complex that traditional data-processing tools are inadequate. Big Data is typically described by three **Vs**: **Volume** (the massive quantity), **Variety** (the many different types), and **Velocity** (the speed at which data is produced and moves).

Today, advanced analytics—especially those powered by **deep learning**—are essential for making sense of Big Data. Computers can be trained via image recognition and natural language processing to detect patterns far more quickly and accurately than humans. The core idea is that **the more you know**, the better you can **predict future outcomes**. As data accumulates and patterns emerge, machines (and humans) can make increasingly informed decisions.

Beyond business, Big Data has transformative applications in fields like **healthcare**, **education**, **disaster prediction**, **emergency response**, **crime prevention**, and **food production**. However, Big Data brings important risks—most notably, **erosion of privacy** and widespread exposure of personal information. In today's world, maintaining true digital privacy is becoming ever more challenging.

In short, Big Data is both a powerful force for innovation and a significant challenge for individual privacy—and it is clearly here to stay.

Concept 7: Data Cleansing

Data cleansing is the process of identifying and correcting—or removing—data that is irrelevant, corrupt, missing, duplicated, or otherwise useless. The goal is to "clean" the dataset so that analytical algorithms can run more efficiently and make more accurate predictions.

Data corruption arises for many reasons, including user mistakes, placeholder or dummy inputs, and workarounds that bypass proper entry. To address this, cleansing typically involves **enhancement**, **harmonization**, and **standardization**: transforming data so it's coherent and consistent across the system.

The process requires locating all faulty, incomplete, or irrelevant items, and then either replacing, deleting, or modifying them so they match other data properly. High-quality data should meet several criteria: it must be valid, accurate, complete, consistent, and uniform.

However, data cleansing has drawbacks. It can be expensive, time-consuming, and pose security risks because data often must be shared for cleaning. Despite these challenges, it's essential—especially in Big Data contexts—for optimizing analytics.

Finally, once data is cleansed, it's important to keep it that way. All new data should conform to the standards and knowledge base already in place. That's why a rigorous data management strategy includes regular cleansing: to catch outdated, incorrect, or inconsistent information over time.

Concept 8: Filling Gaps in Data

Data gaps occur when certain fields in a dataset contain no information. These missing values slow down algorithmic processing and may eliminate important insights that organizations rely on for decision-making. Because missing data can significantly affect performance, various strategies are used to fill these gaps, though no single method is universally correct.

Common heuristic techniques include **carrying forward known values** from similar data—such as inserting a missing postal code based on the city— or **using historical data** from a comparable time period, like the same month in a previous year. Another method is filling gaps using **average values** of similar data points, though this approach can be risky, as it may reinforce existing assumptions rather than reveal new trends. Finally, when dealing with extremely large datasets, a practical option may be to **delete records** that contain missing fields, since the loss of a small number of entries is unlikely to affect overall analysis.

Overall, handling data gaps is essential to ensure efficient computation and maintain the reliability of analytical outcomes.

Concept 9: A Fast Snapshot of Machine Learning