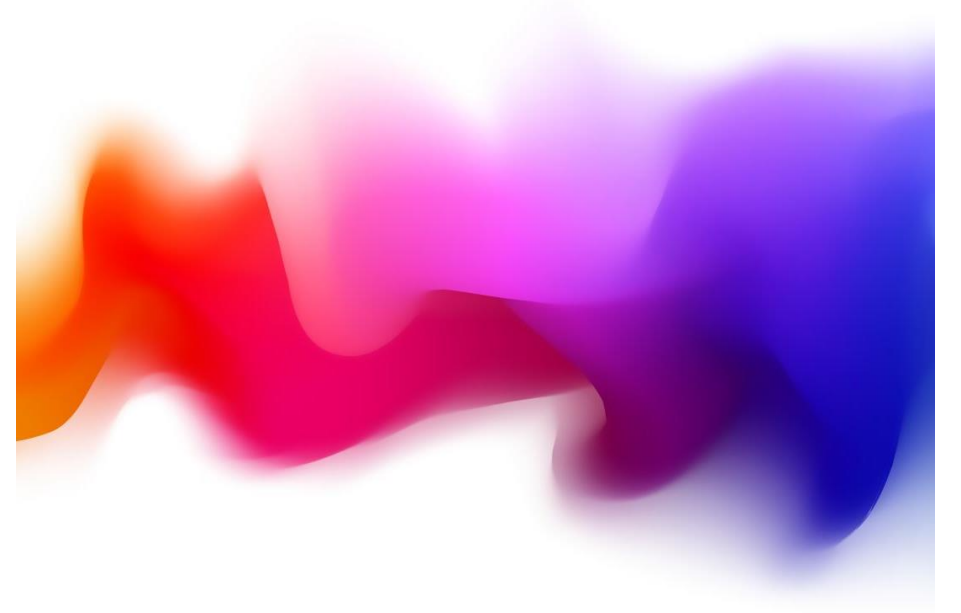


Foundation of Data Science – Topic 3

Anitya Gupta

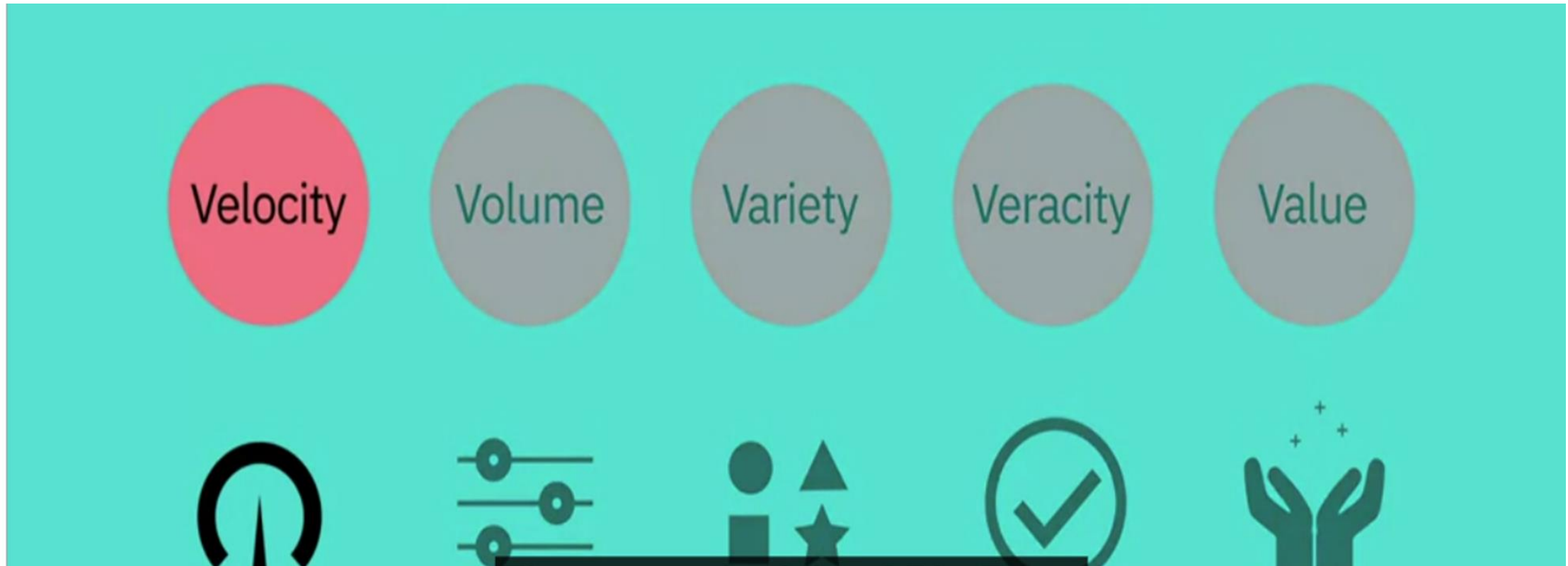


Foundation of Big Data

Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines.

Requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.

The V's of Big Data



1. **Velocity:** speed at which data accumulates. Data is being generated extremely fast, in a process that never stops.
2. **Volume:** scale of the data, or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.
3. **Variety:** diversity of the data. Also accumulates the data comes from different sources, machines, people, and processes, both internal and external to organizations.
4. **Veracity:** quality and origin of data, and its conformity to facts and accuracy. Attributes include consistency, completeness, integrity, and ambiguity.
5. **Value:** Our ability and need to turn data into value. Value isn't just profit. It may have medical or social benefits, as well as customer, employee, or personal satisfaction. The main reason that people invest time to understand Big Data is to derive value from it.

Hadoop

- A. The scientists basically sliced the files into pieces and then send to thousands of computers in cluster.
- B. Each computer run the piece of the file and program and send back the result.
- C. The results are then sorted and send back to other process.
- D. 1st Process called “Map” and 2nd Process called “Reduce”.